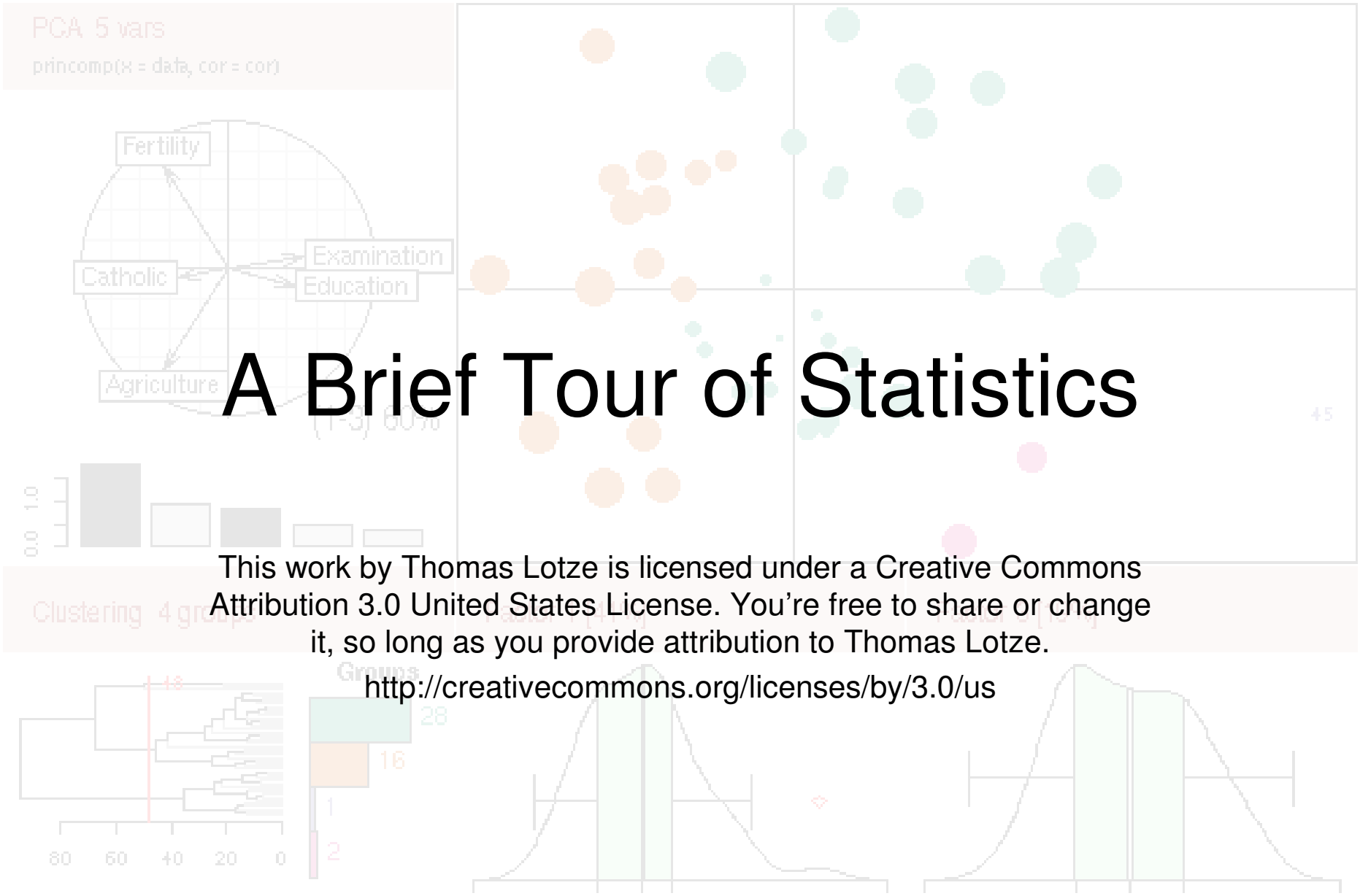
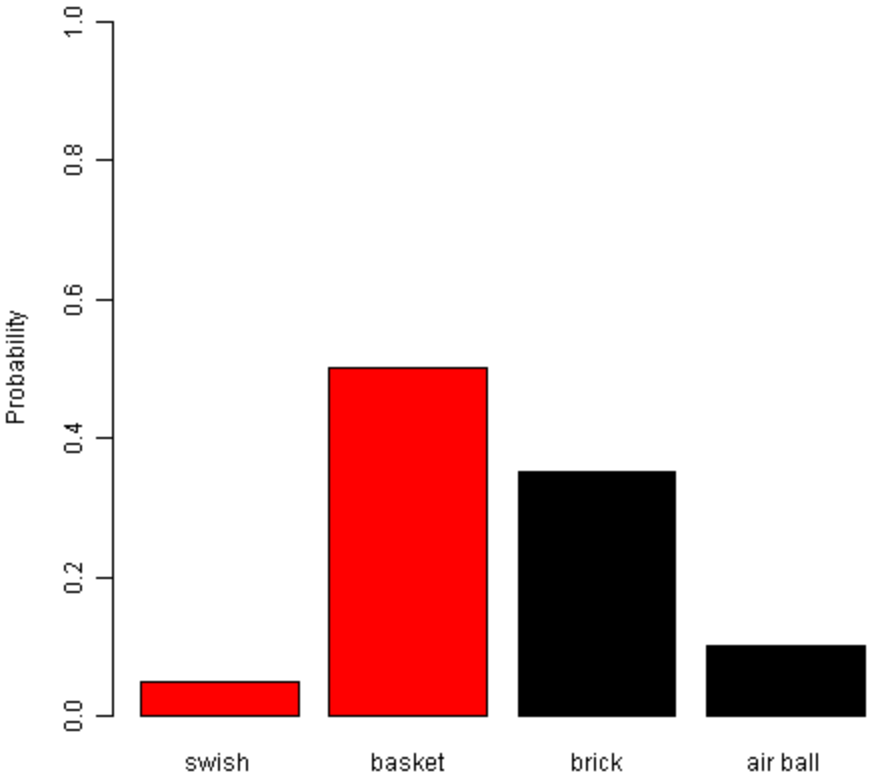
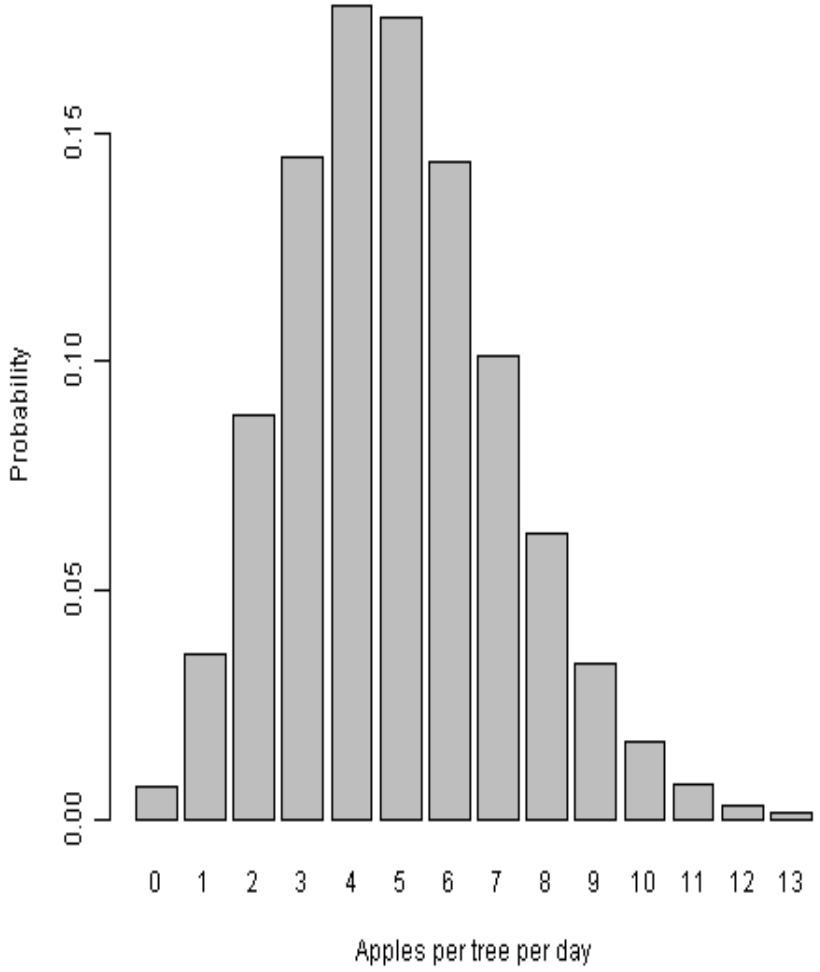
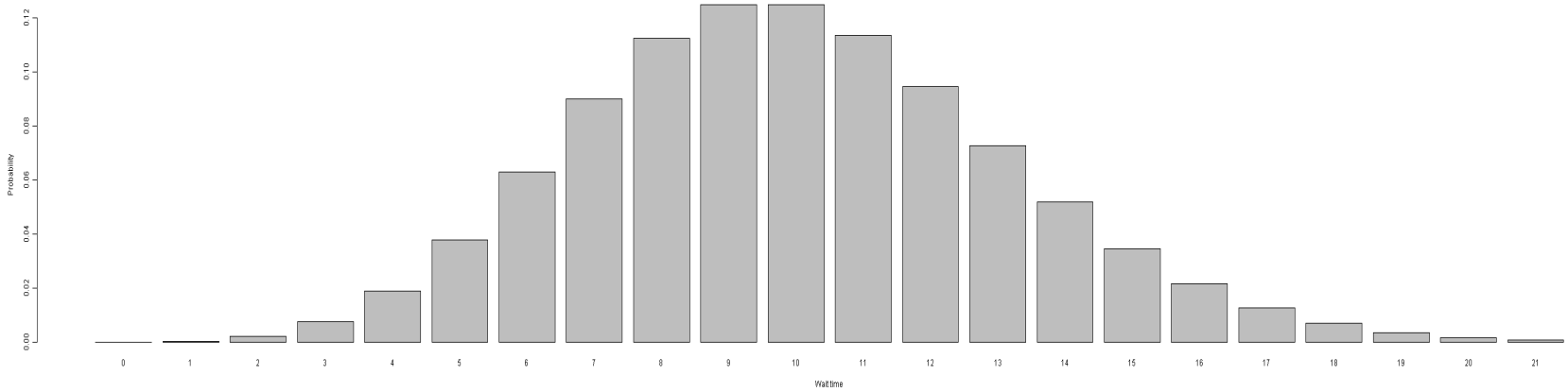


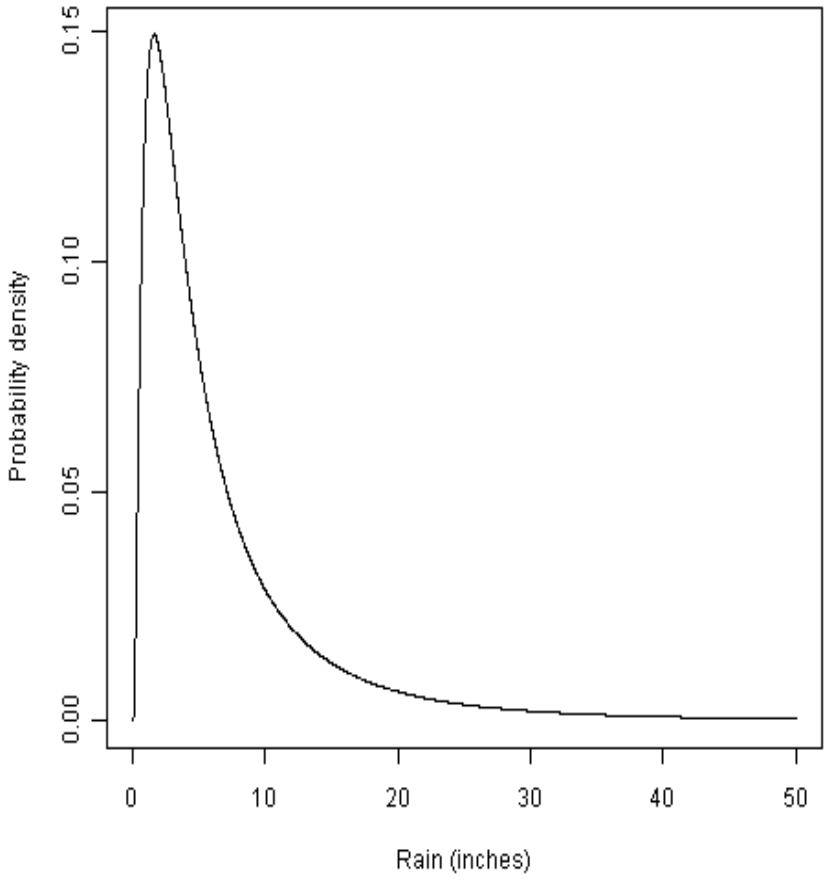
A Brief Tour of Statistics

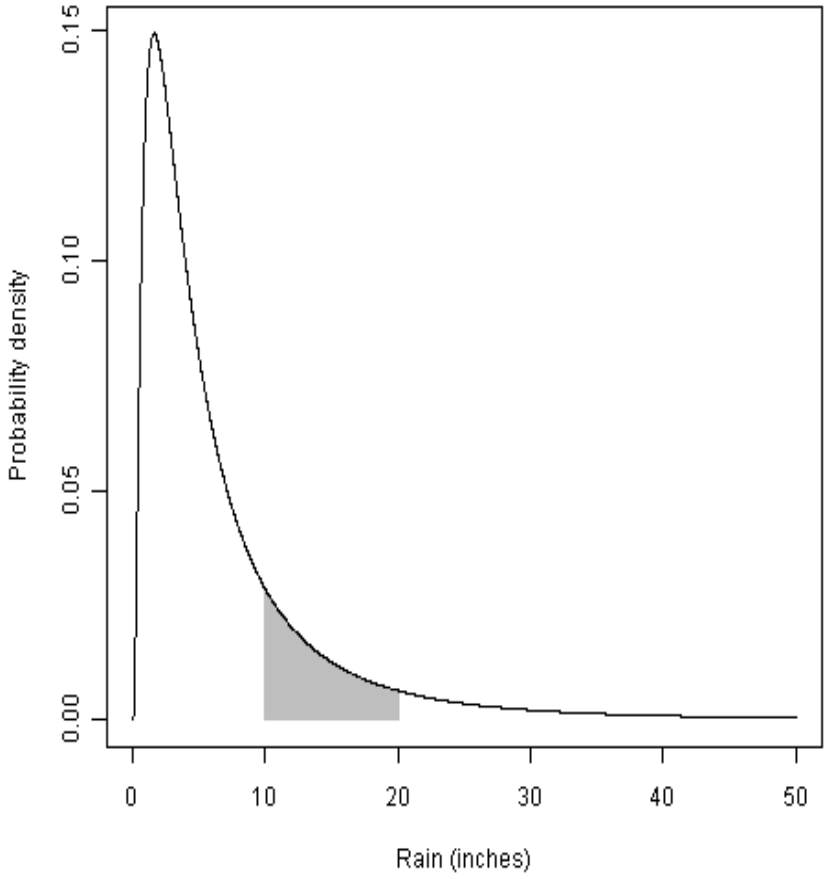
- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- Distribution Parameters
- Estimating Parameters
- Confidence Intervals and P-values
- Why Gaussian?
- Regression
- Logistic Regression
- Why Not Gaussian?
- Bootstrapping
- Multiple Testing
- Useful Tools



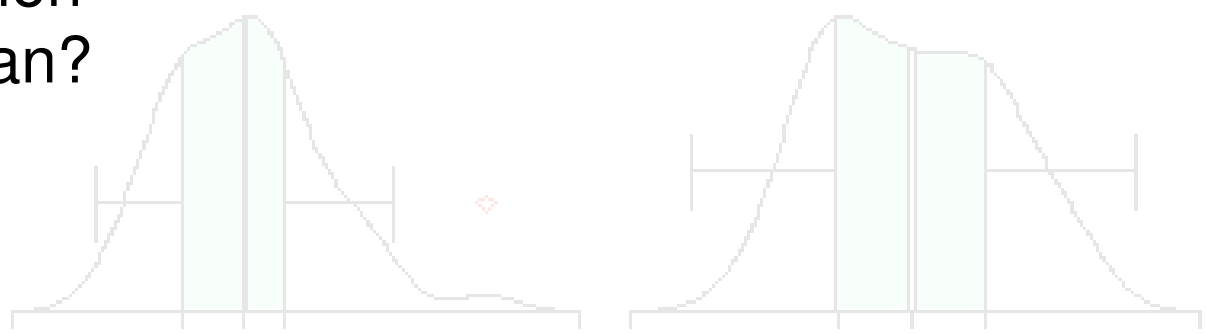




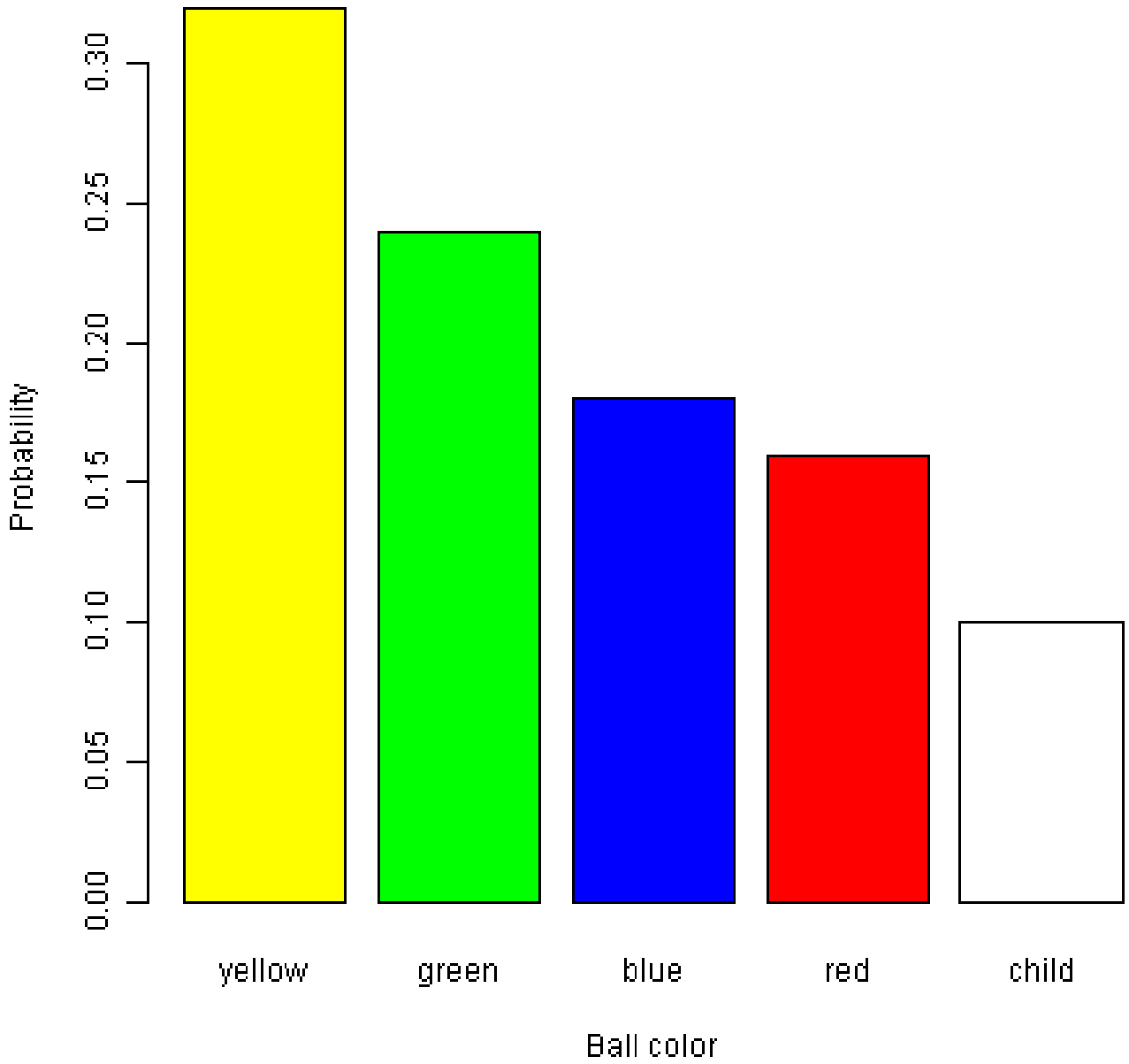


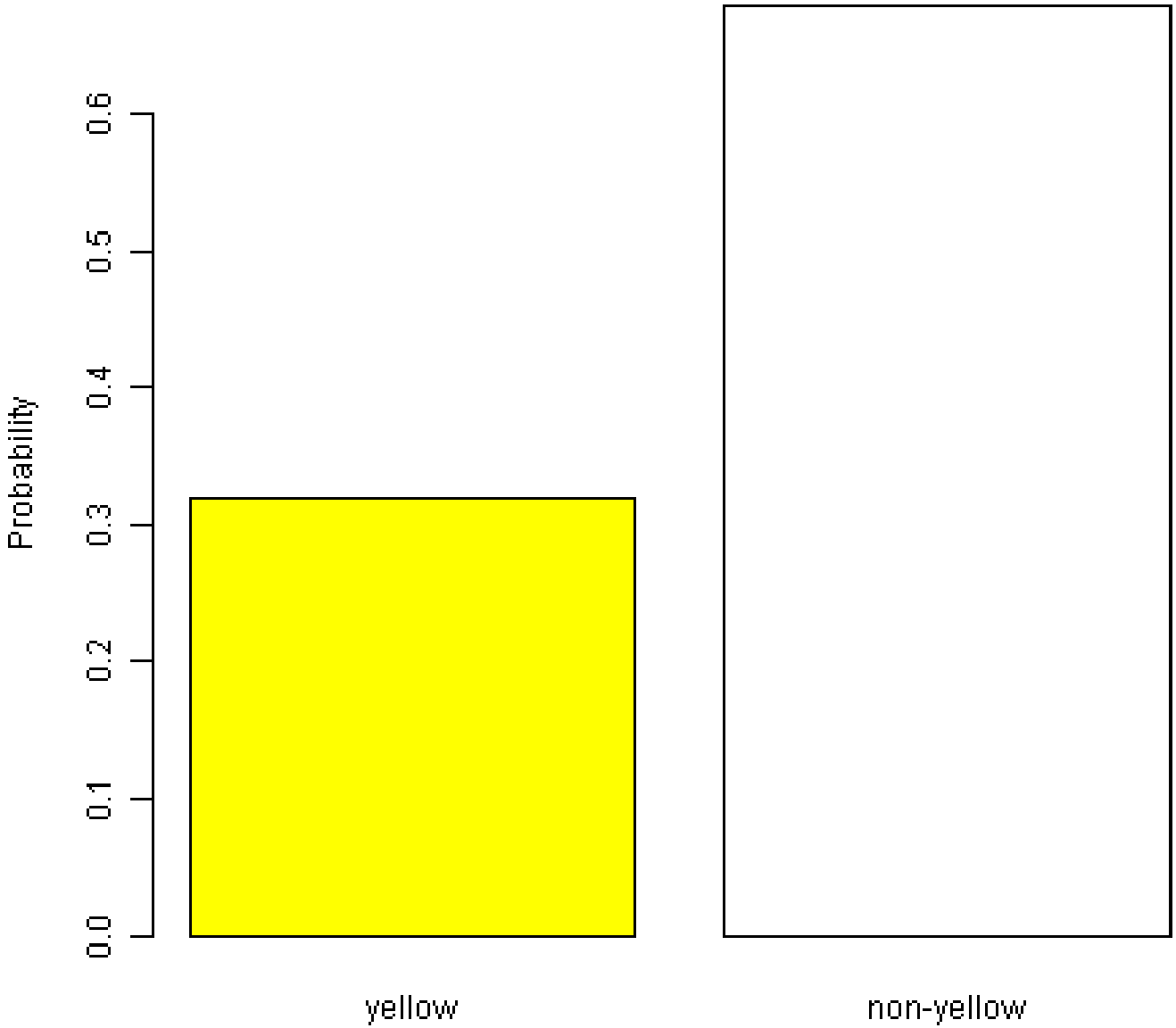


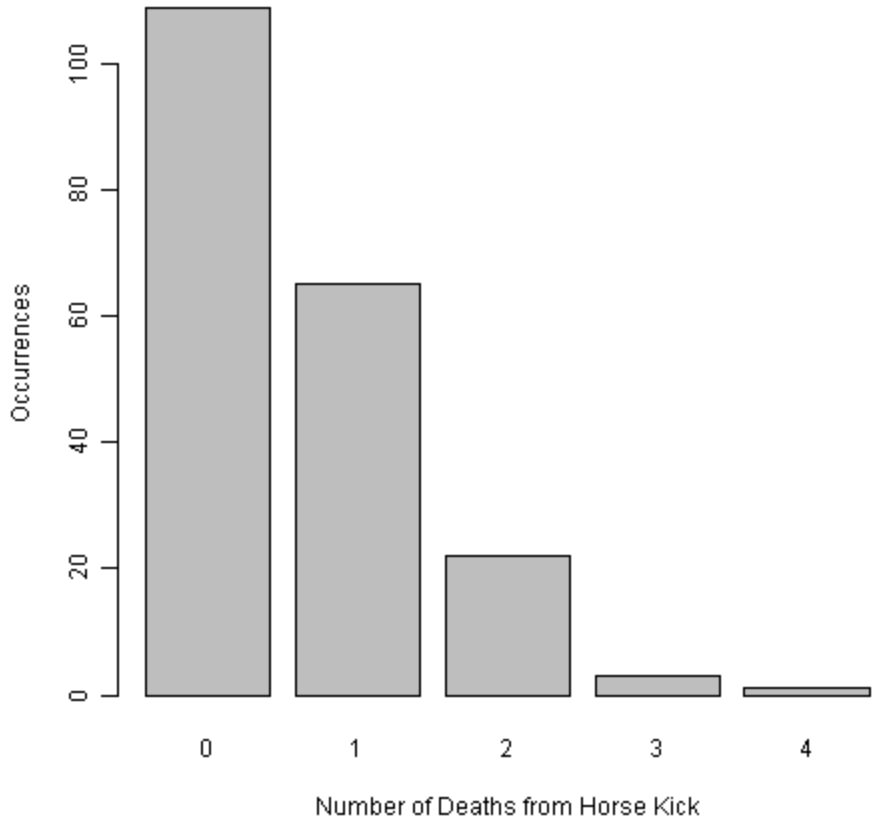
- What are Distributions?
- **Models**
  - **Binomial**
  - **Poisson**
  - **Uniform**
  - **Gaussian (normal)**
- Using Distributions to Answer Questions
- Distribution Parameters
- Estimating Parameters
- Confidence Intervals and P-values
- Why Gaussian?
- Regression
- Logistic Regression
- Why Not Gaussian?
- Bootstrapping
- Multiple Testing
- Useful Tools

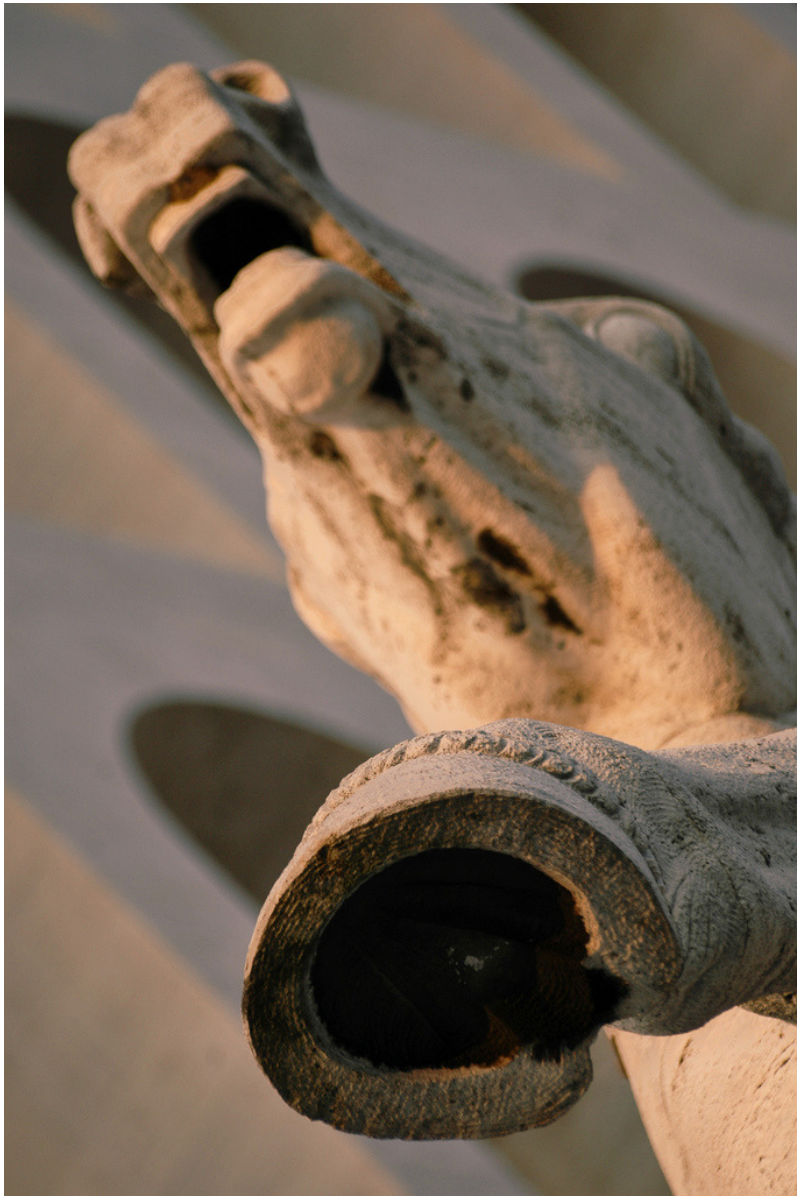


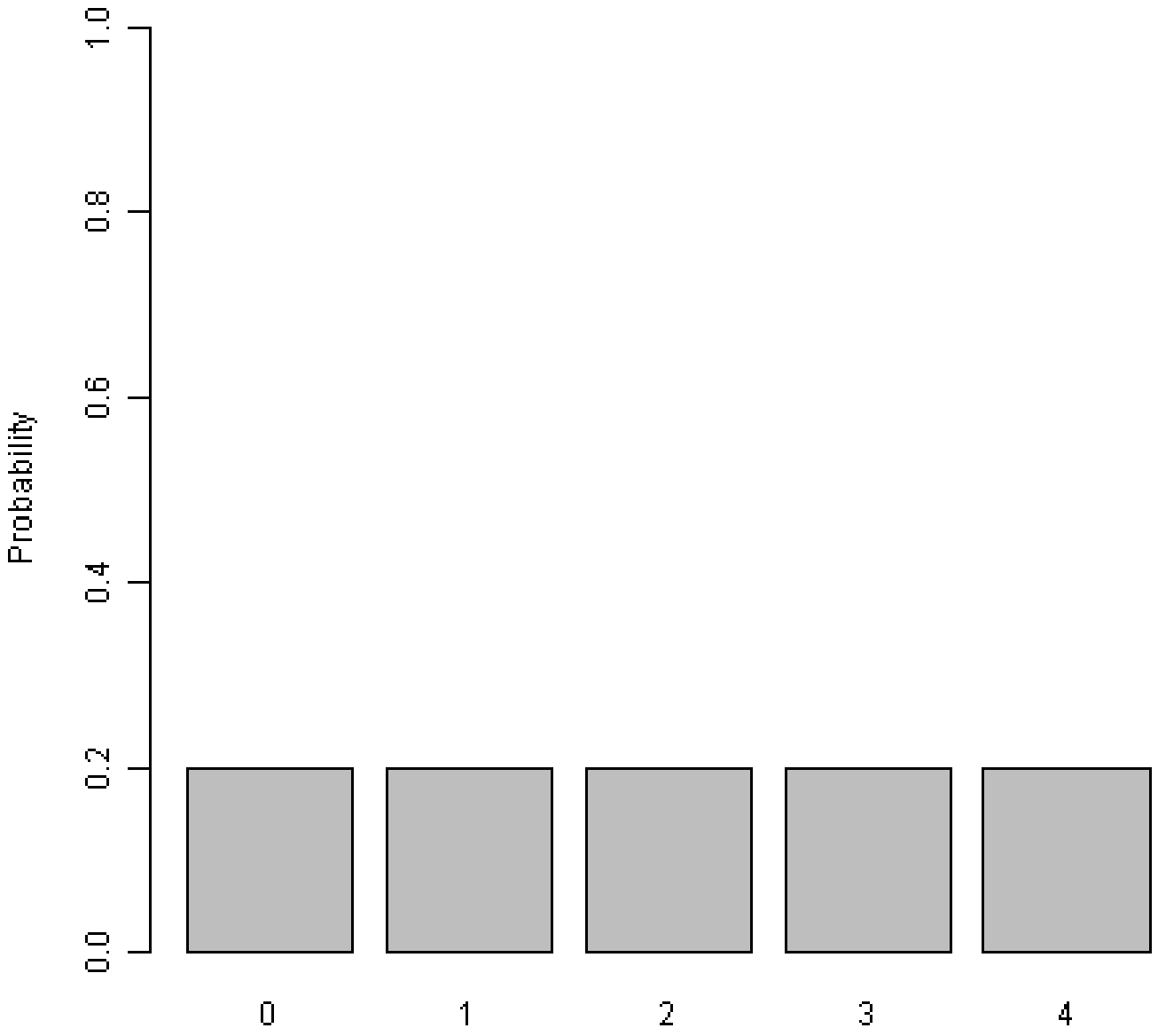
Creative Commons licensed, from Jason Tromm on flickr

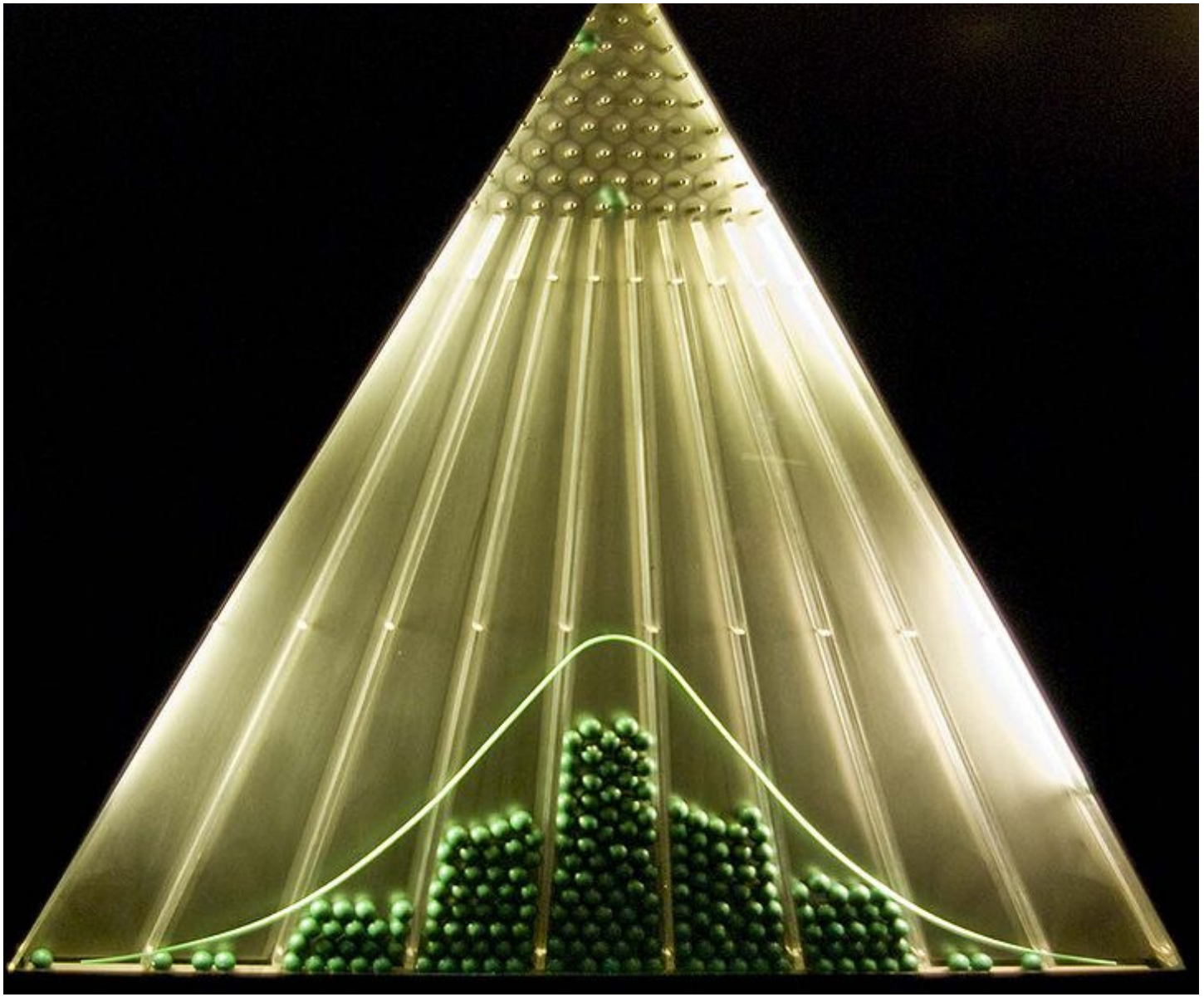

Probability
0.00
0.05
0.10
0.15
0.20
0.25
0.30
yellow
green
blue
red
child
Ball color


Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
yellow
non-yellow
Ball color

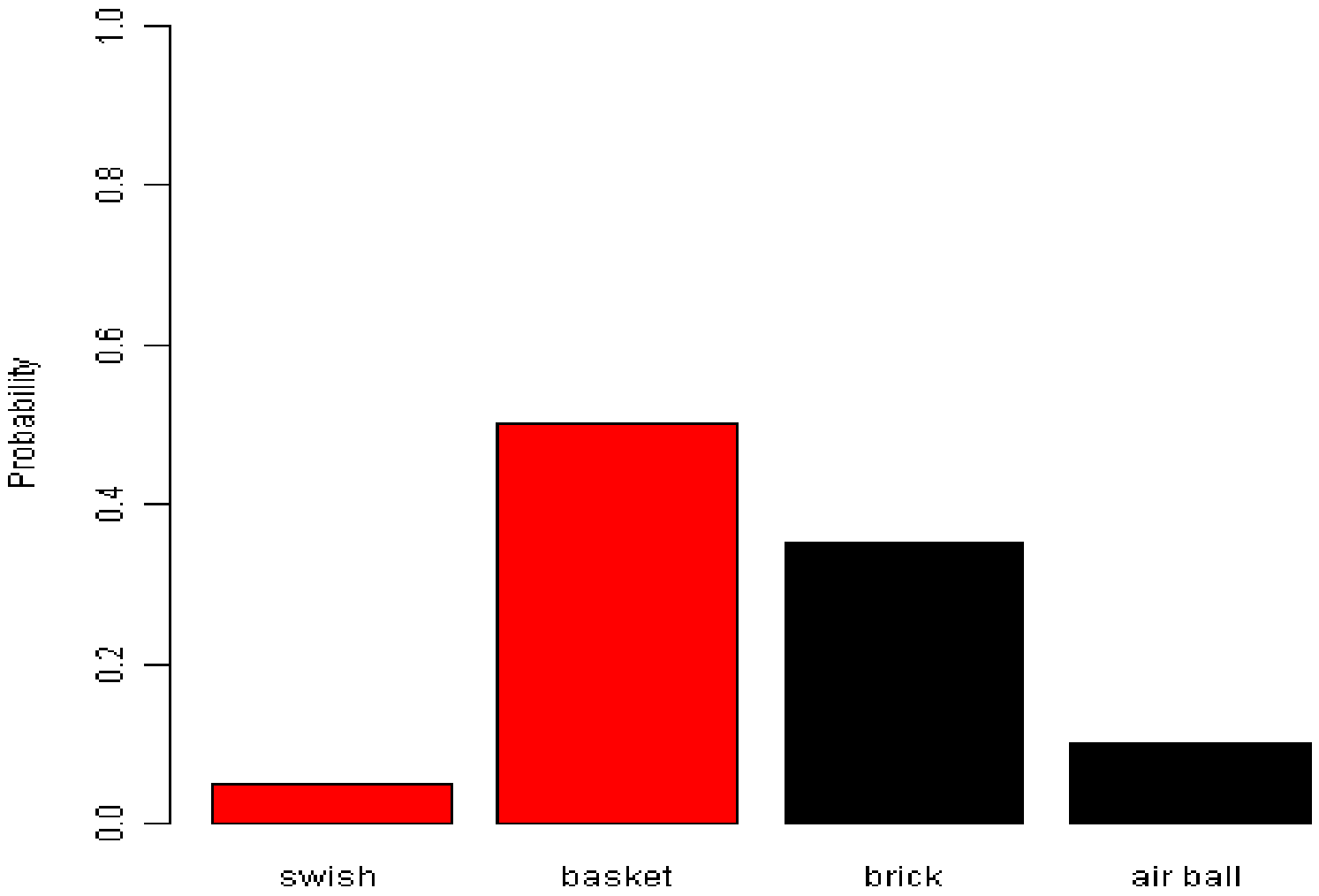


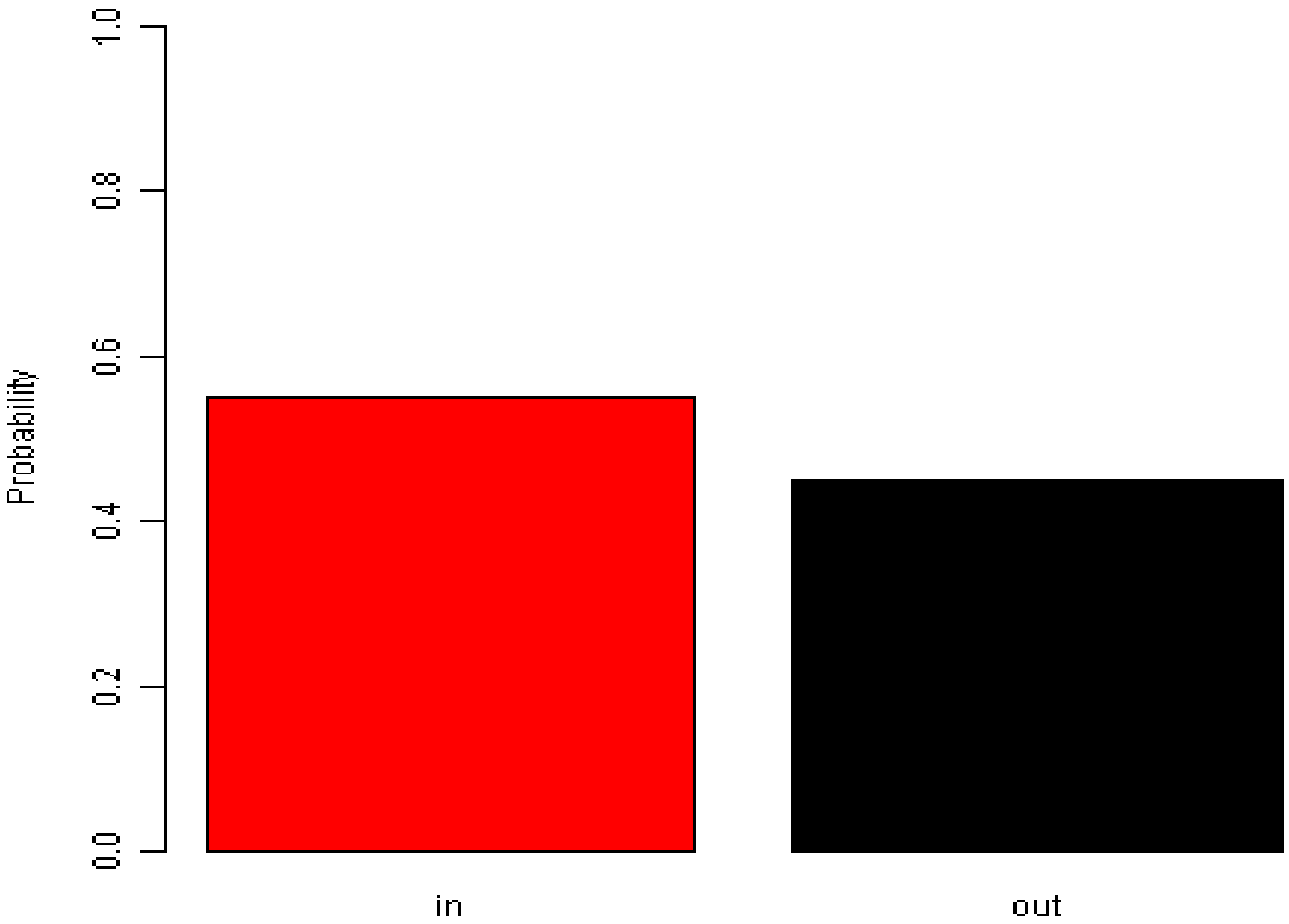


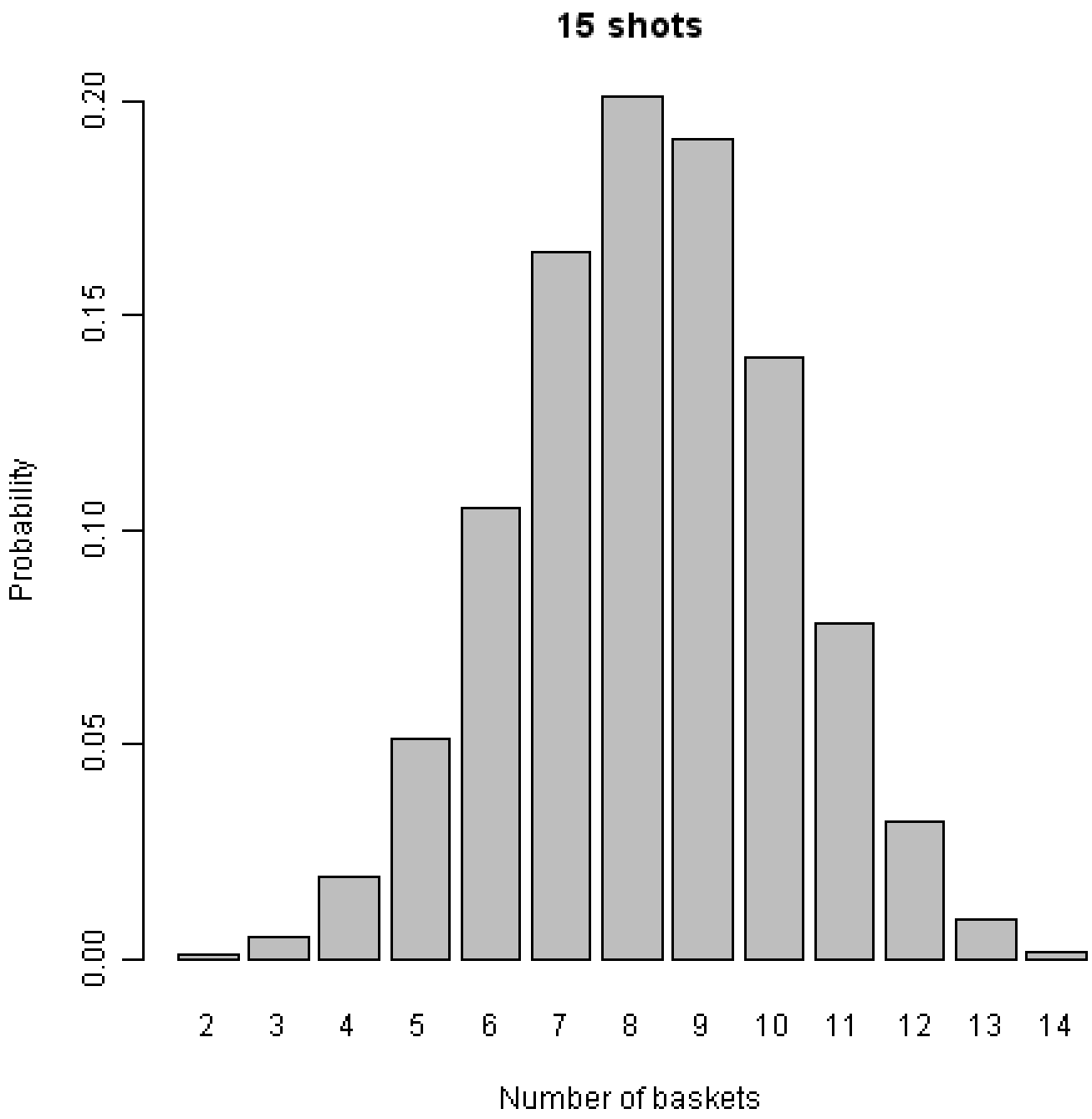

Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4

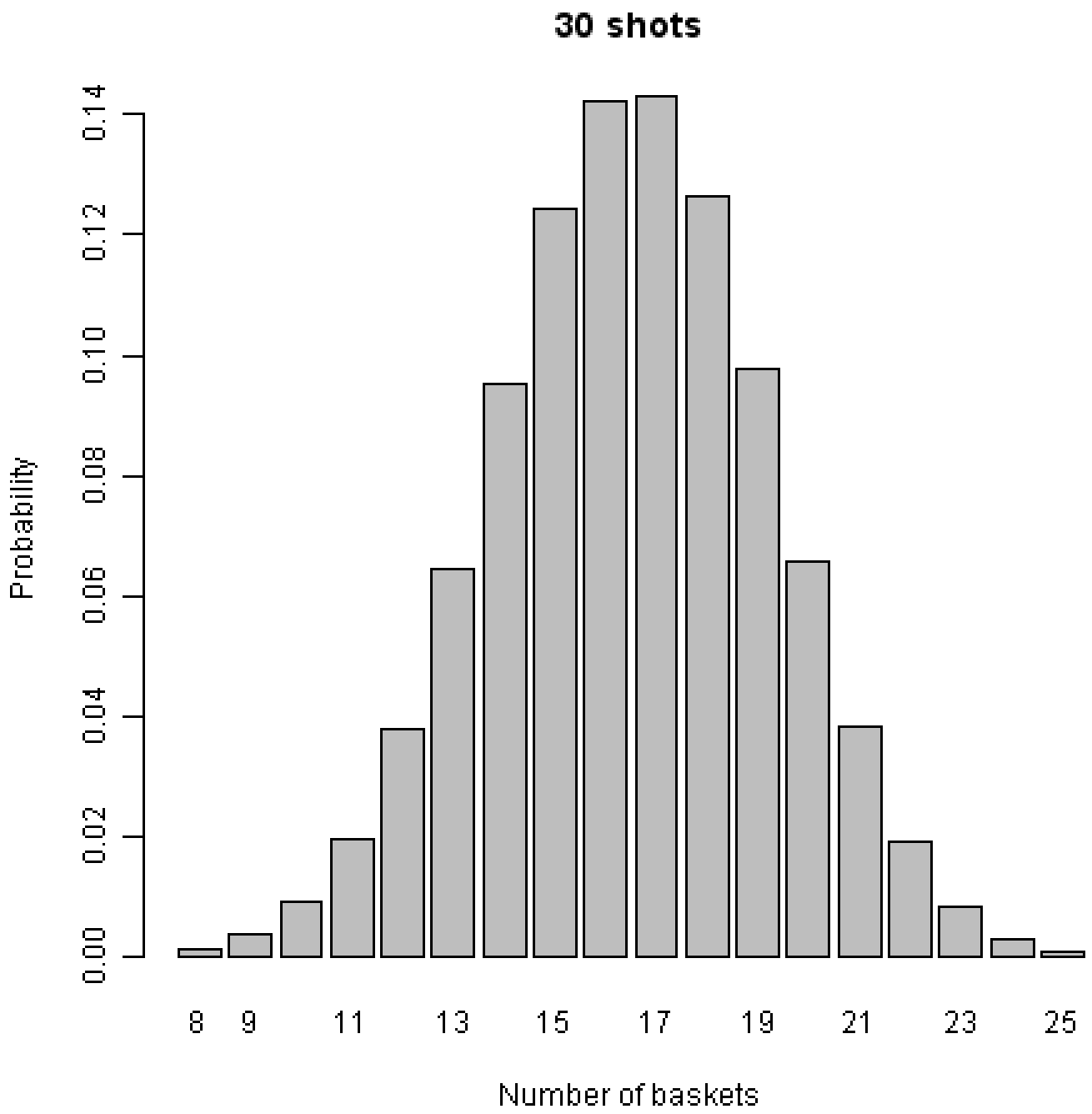


- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- **Using Distributions to Answer Questions**
- Distribution Parameters
- Estimating Parameters
- Confidence Intervals and P-values
- Why Gaussian?
- Regression
- Logistic Regression
- Why Not Gaussian?
- Bootstrapping
- Multiple Testing
- Useful Tools


Probability
0.0
0.2
0.4
0.6
0.8
1.0
swish
basket
brick
air ball

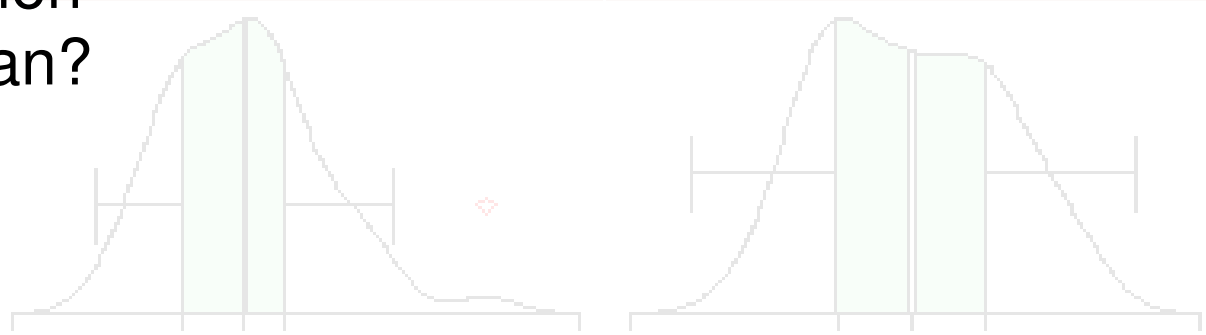

Probability
0.0
0.2
0.4
0.6
0.8
1.0
in
out


15 shots
Probability
0.00
0.05
0.10
0.15
0.20
2
3
4
5
6
7
8
9
10
11
12
13
14
Number of baskets


30 shots
Probability
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
8
9
11
13
15
17
19
21
23
25
Number of baskets

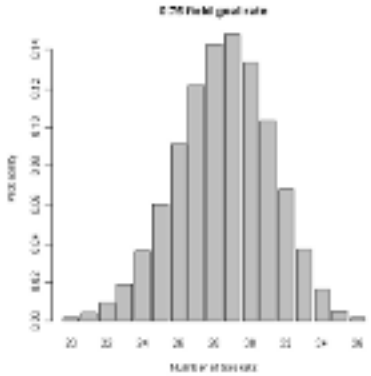
- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- **Distribution Parameters**
- Estimating Parameters
- Confidence Intervals and P-values
- Why Gaussian?
- Regression
- Logistic Regression
- Why Not Gaussian?
- Bootstrapping
- Multiple Testing
- Useful Tools

- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- Distribution Parameters
- **Estimating Parameters**
- Confidence Intervals and P-values
- Why Gaussian?
- Regression
- Logistic Regression
- Why Not Gaussian?
- Bootstrapping
- Multiple Testing
- Useful Tools



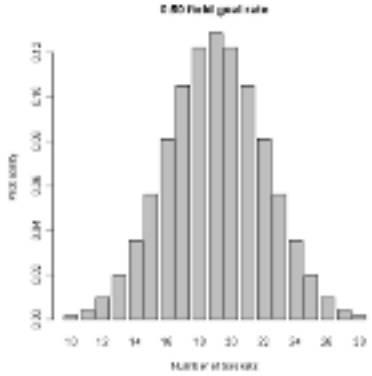
21/38

p=0.75



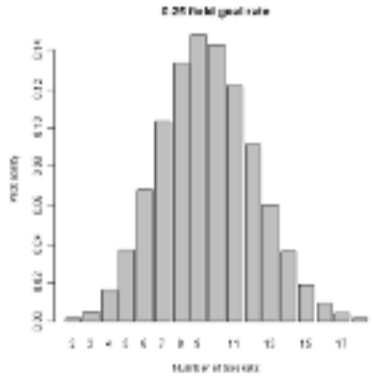
20,28,31,25,29/38

p=0.5

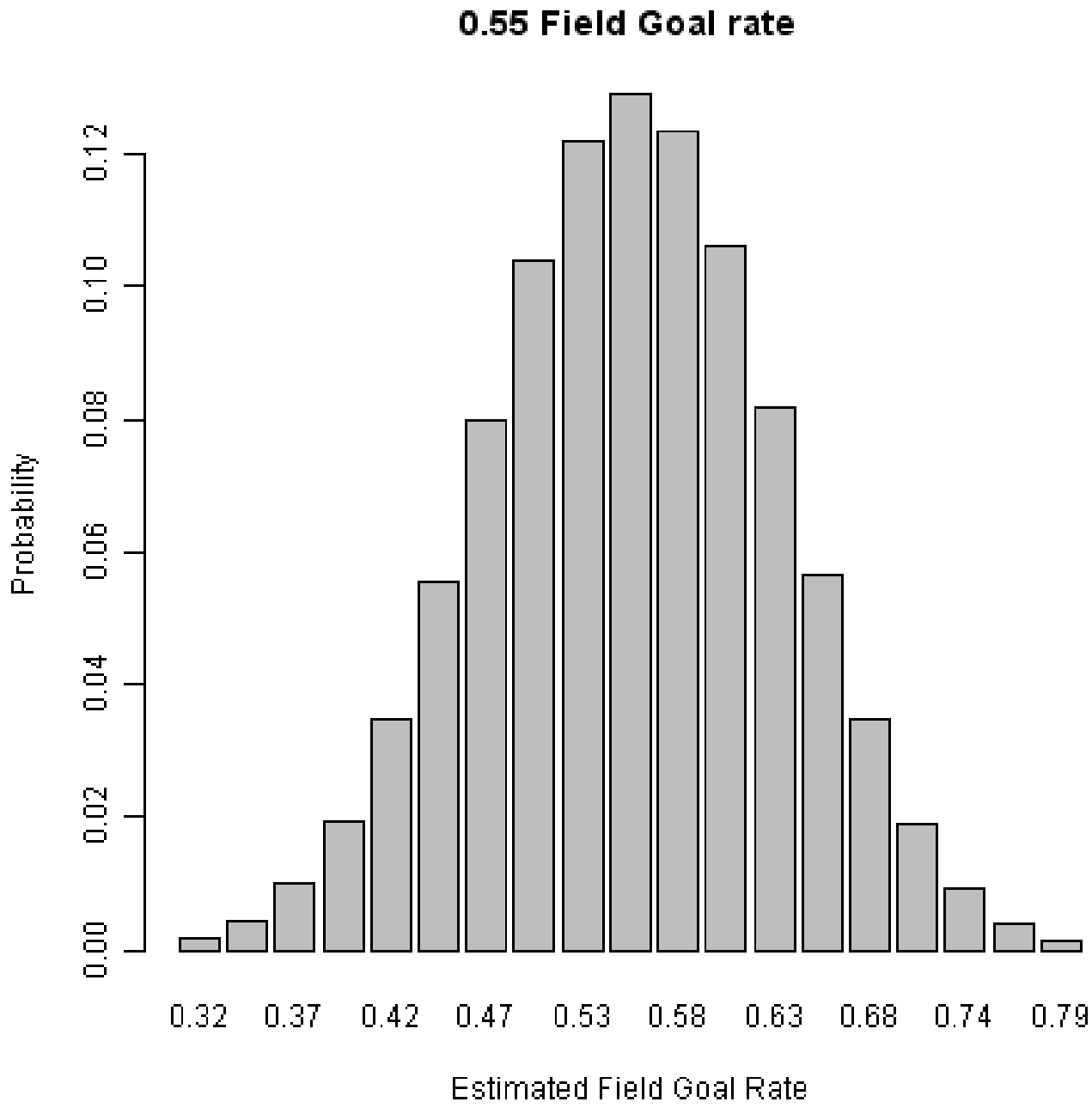


22,15,21,18,18/38

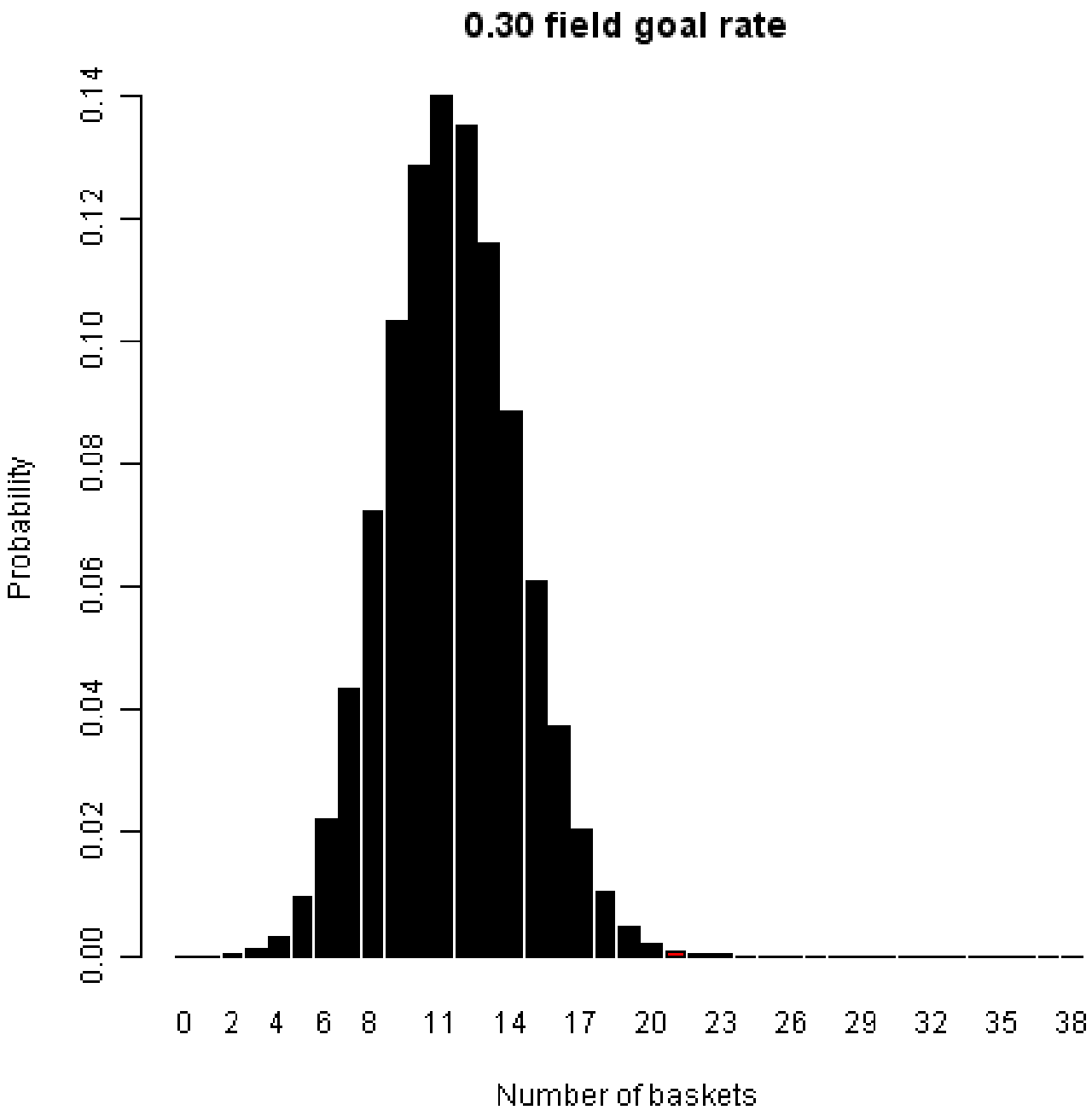
p=0.25

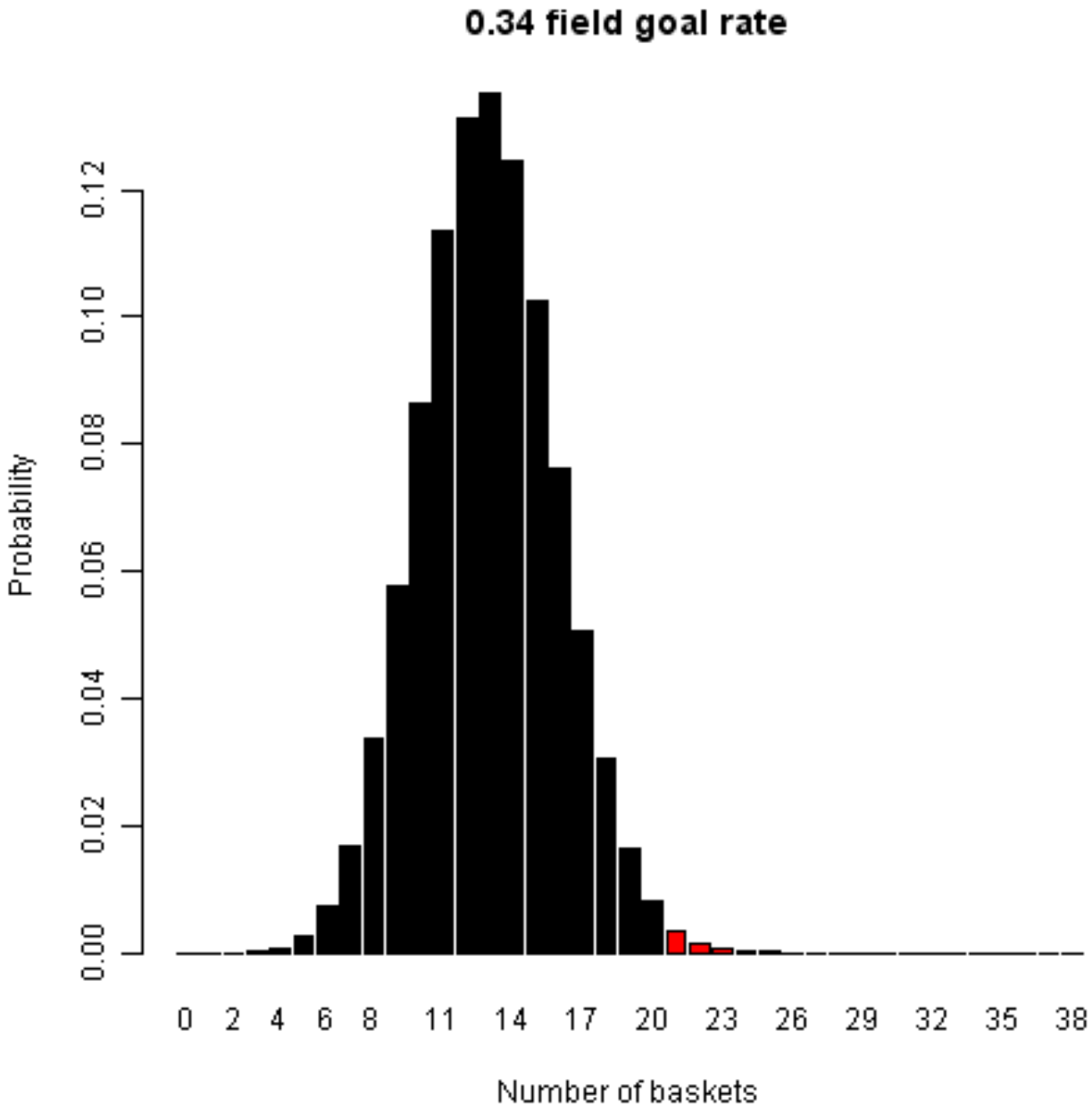


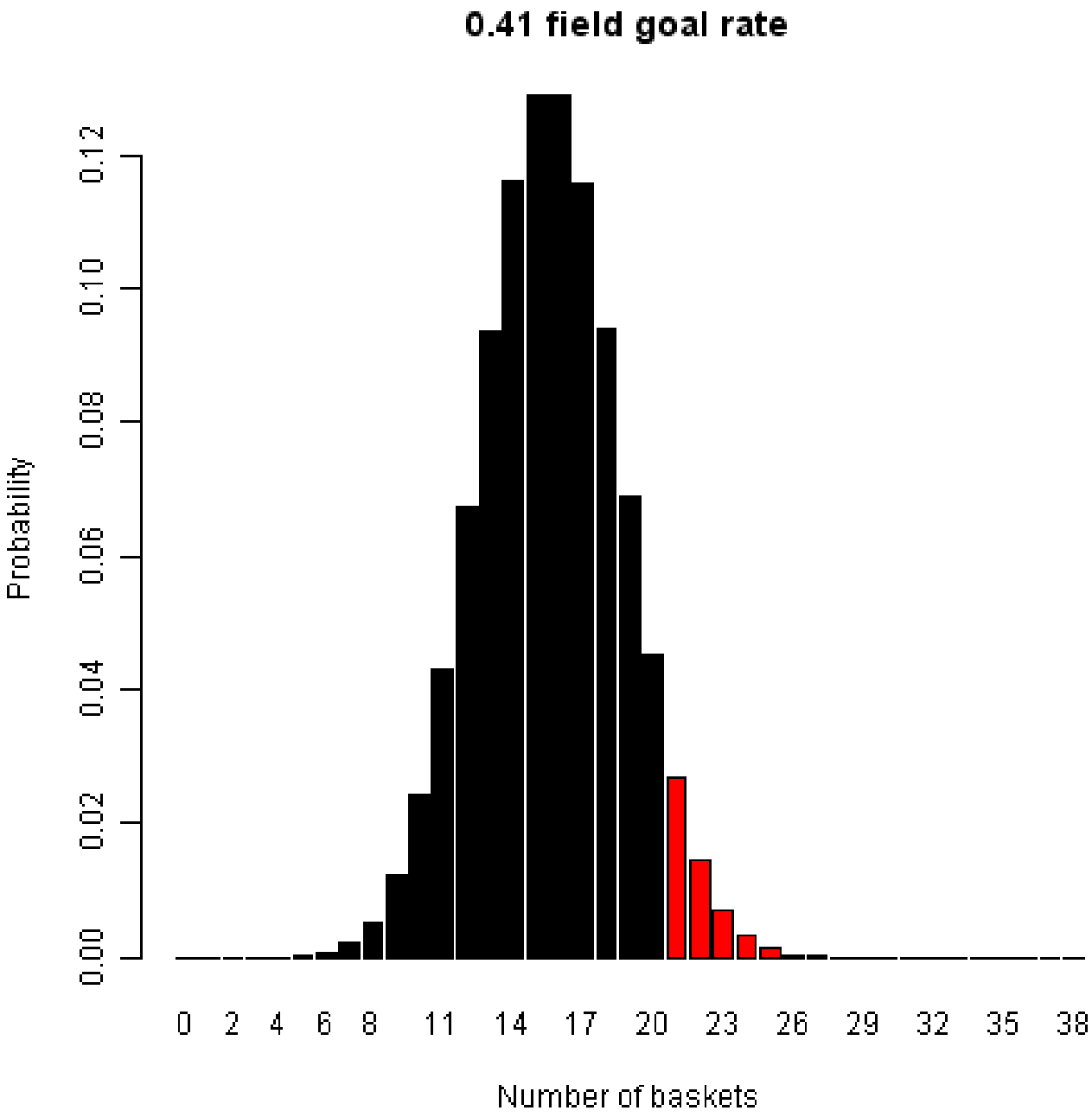
12,10,11,11,7/38

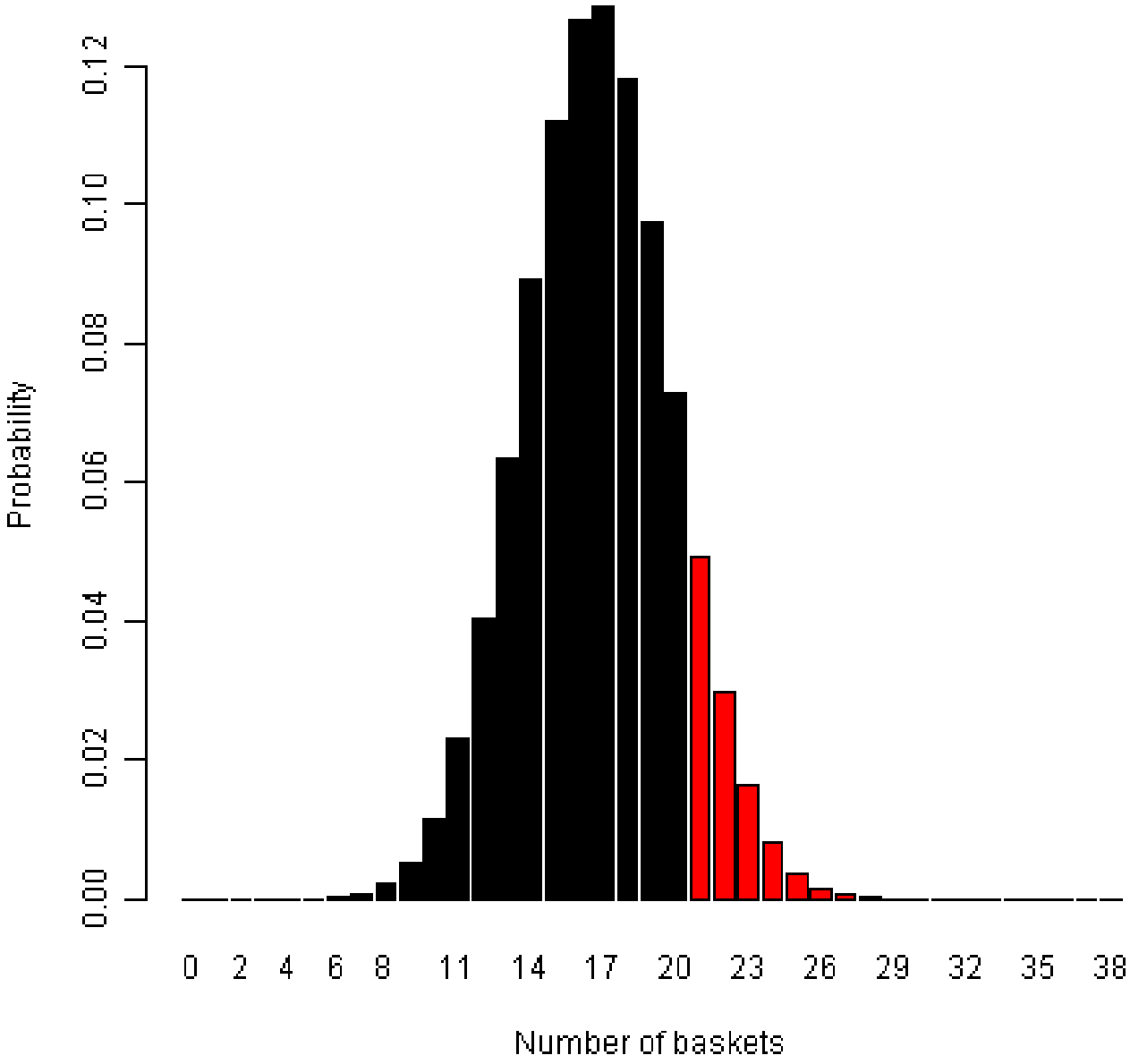

0.55 Field Goal rate
Probability
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.32
0.37
0.42
0.47
0.53
0.58
0.63
0.68
0.74
0.79
Estimated Field Goal Rate

- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- Distribution Parameters
- Estimating Parameters
- **Confidence Intervals and P-values**
- Why Gaussian?
- Regression
- Logistic Regression
- Why Not Gaussian?
- Bootstrapping
- Multiple Testing
- Useful Tools


0.30 field goal rate
Probability
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
2
4
6
8
11
14
17
20
23
26
29
32
35
38
Number of baskets

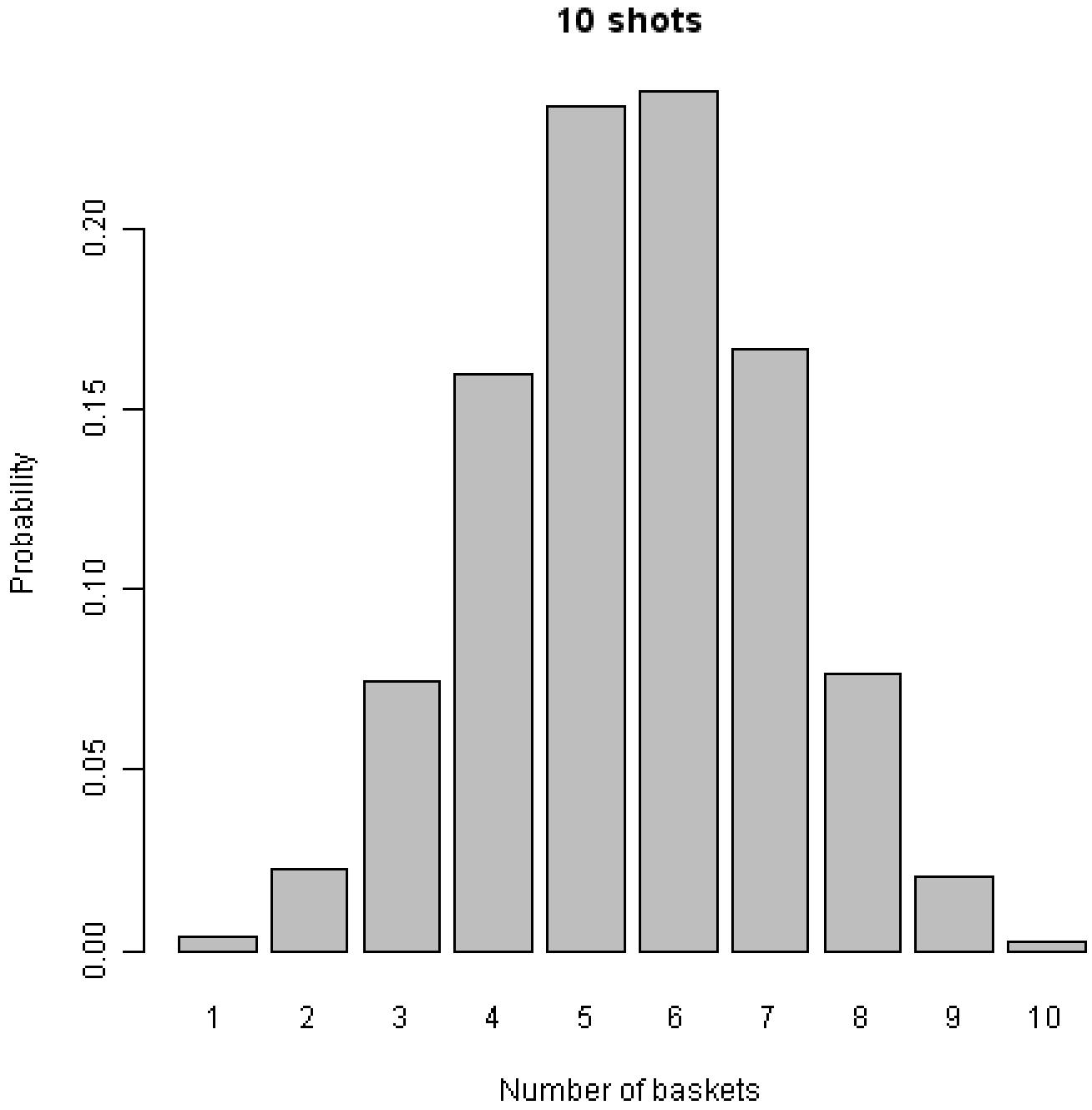

0.34 field goal rate
Probability
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
2
4
6
8
11
14
17
20
23
26
29
32
35
38
Number of baskets

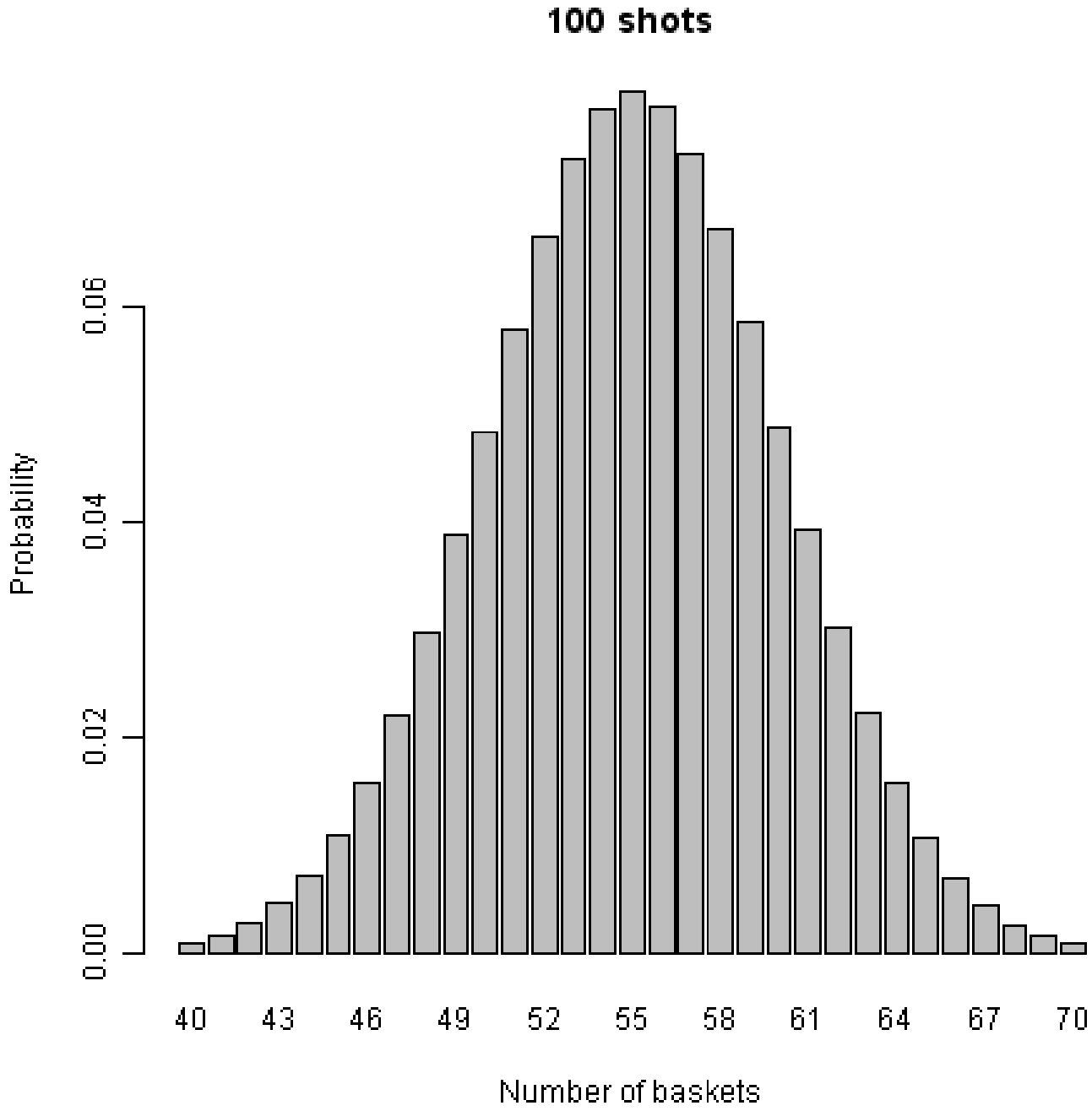

0.41 field goal rate
Probability
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
2
4
6
8
11
14
17
20
23
26
29
32
35
38
Number of baskets

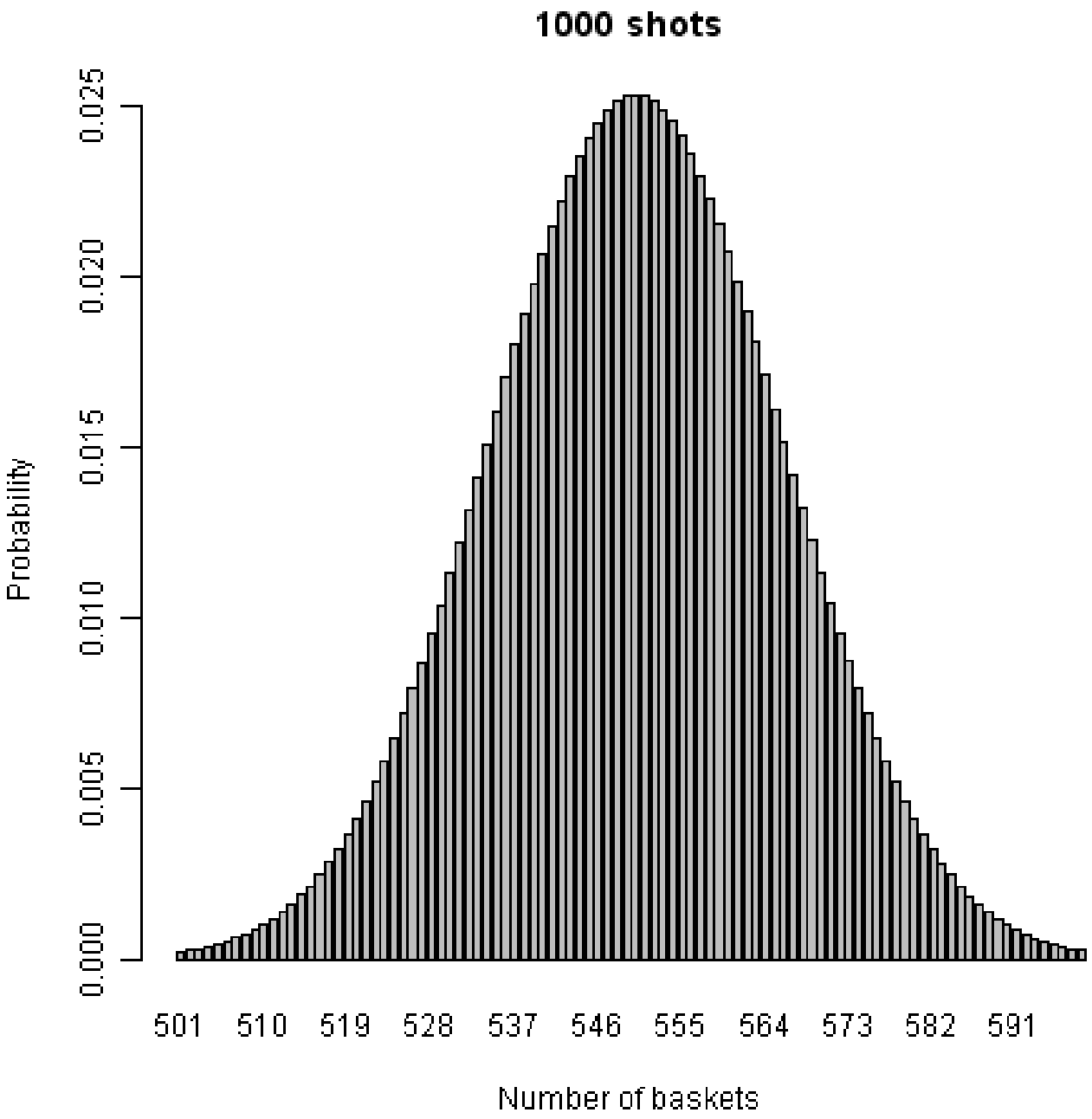

0.44 field goal rate
Probability
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
2
4
6
8
11
14
17
20
23
26
29
32
35
38
Number of baskets

Actual Field Goal Rate	Probability of 21 or more successes
0.30	0.001
0.34	0.01
0.41	0.05
0.44	0.1

- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- Distribution Parameters
- Estimating Parameters
- Confidence Intervals and P-values
- **Why Gaussian?**
- Regression
- Logistic Regression
- Why Not Gaussian?
- Bootstrapping
- Multiple Testing
- Useful Tools


10 shots
Probability
0.00
0.05
0.10
0.15
0.20
1
2
3
4
5
6
7
8
9
10
Number of baskets

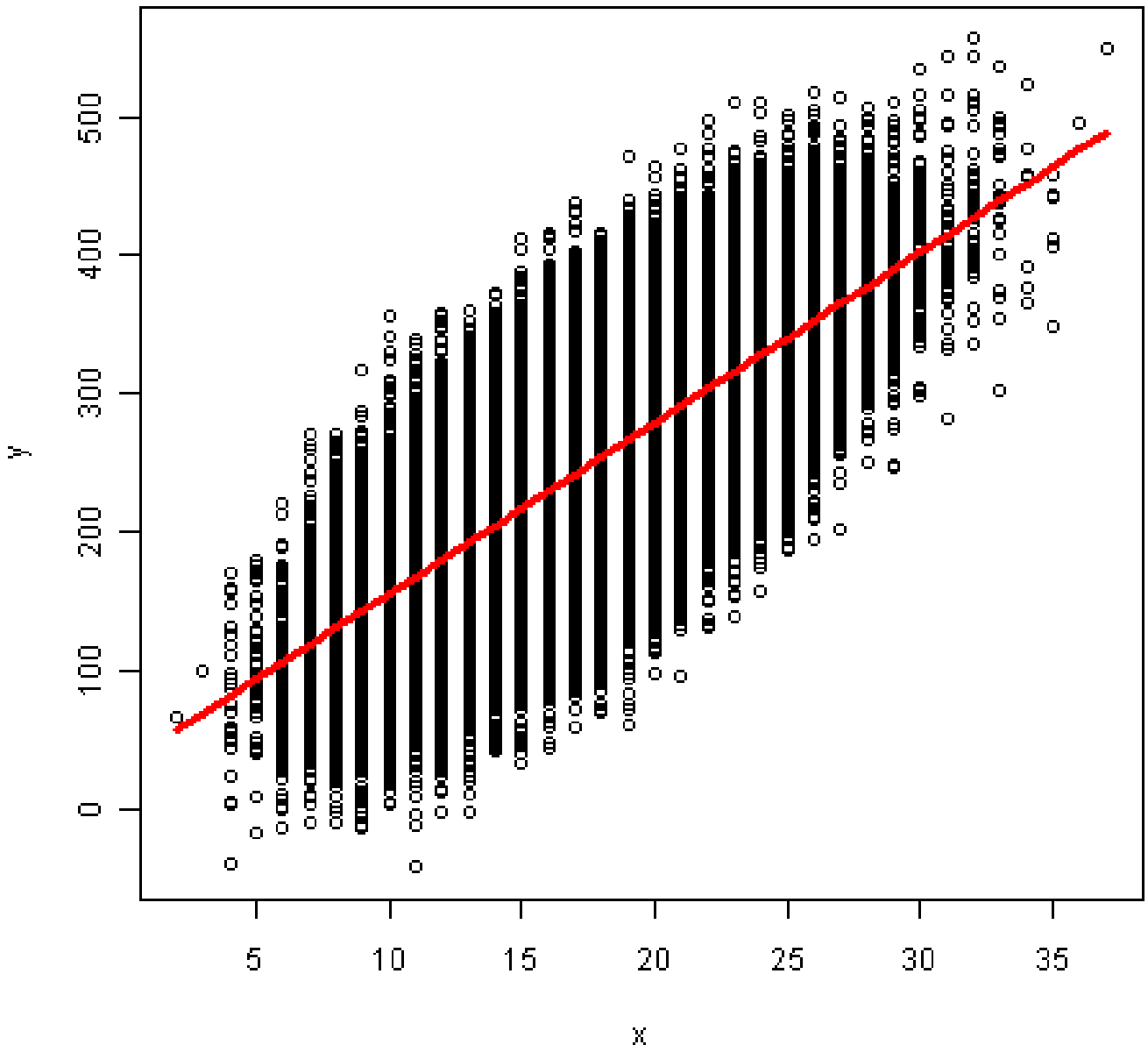

100 shots
Probability
0.00
0.02
0.04
0.06
40
43
46
49
52
55
58
61
64
67
70
Number of baskets

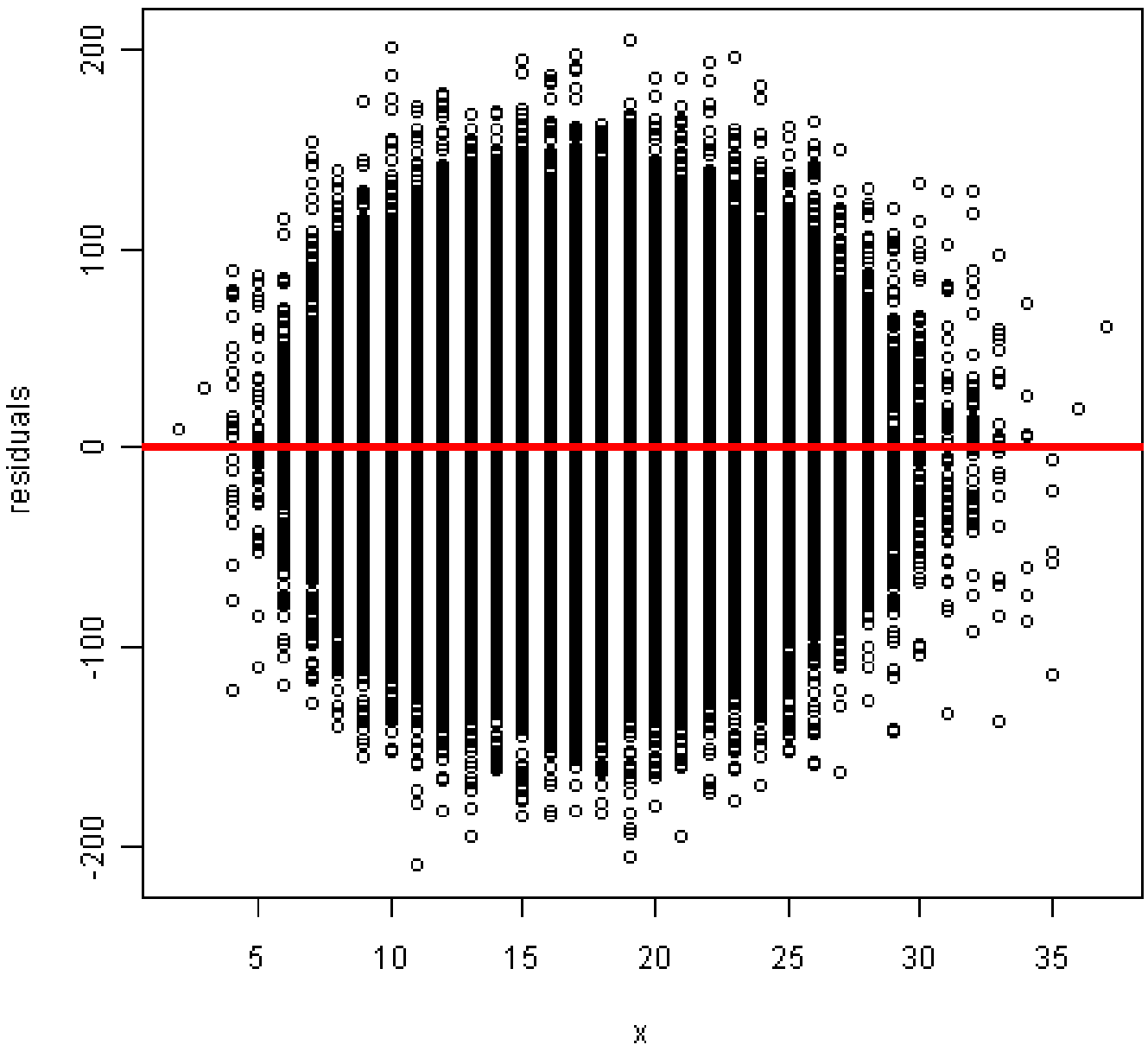

1000 shots
Probability
0.000
0.005
0.010
0.015
0.020
0.025
501
510
519
528
537
546
555
564
573
582
591
Number of baskets

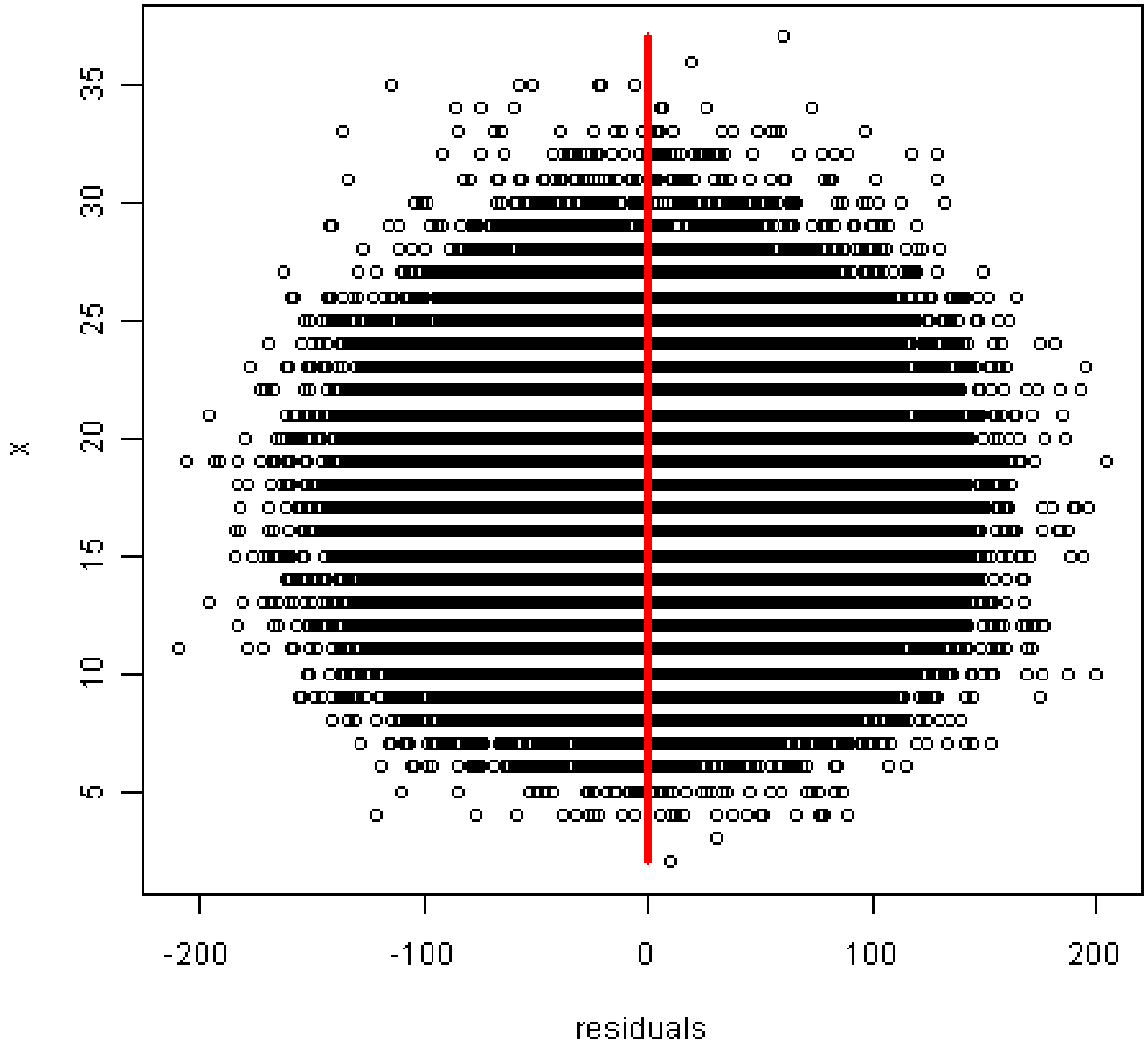
- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- Distribution Parameters
- Estimating Parameters
- Confidence Intervals and P-values
- Why Gaussian?
- **Regression**
- Logistic Regression
- Why Not Gaussian?
- Bootstrapping
- Multiple Testing
- Useful Tools

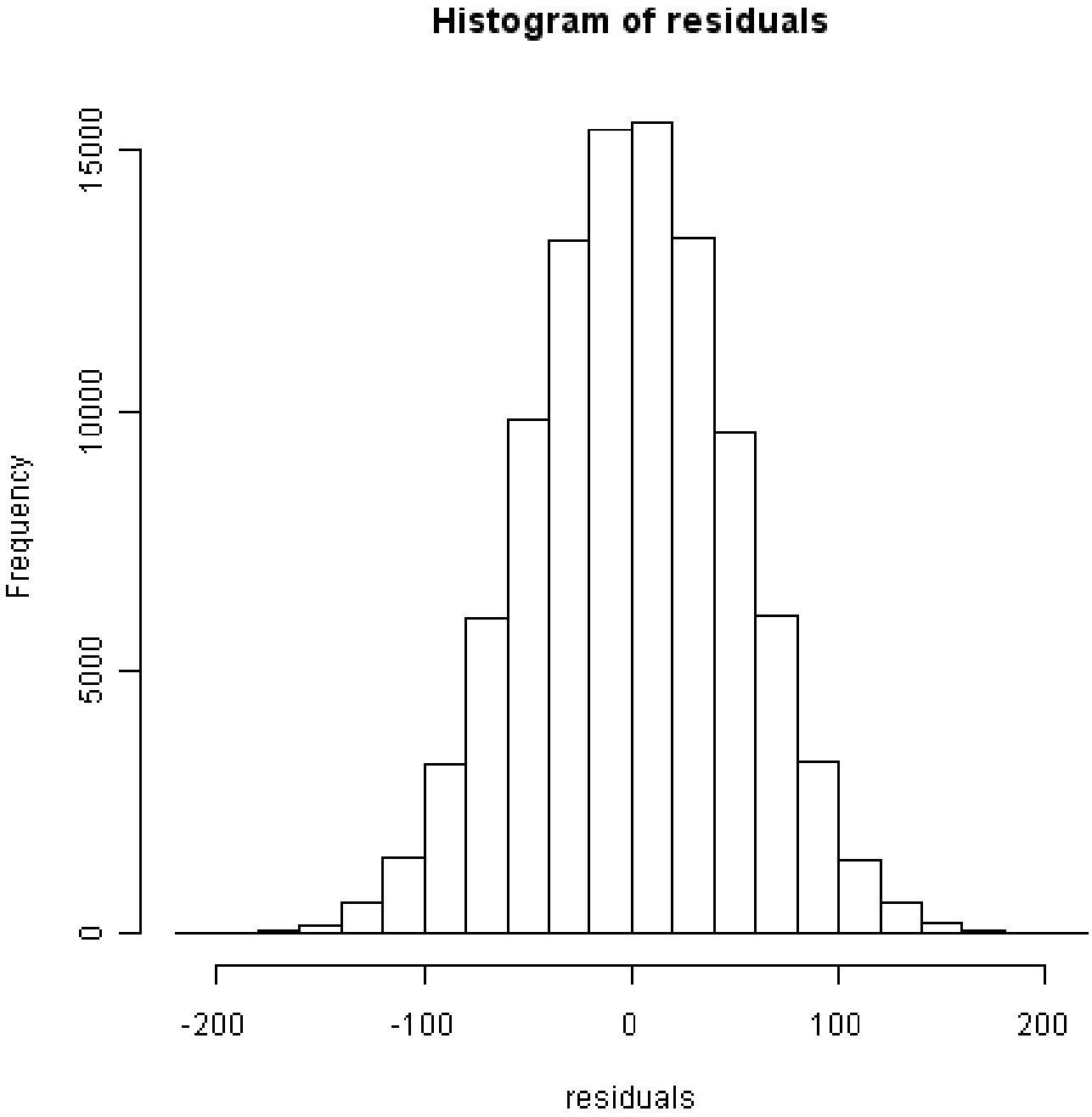
$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$


y
0
100
200
300
400
500
5
10
15
20
25
30
35
x


residuals
x
200
100
0
-100
-200
5
10
15
20
25
30
35


x
5
10
15
20
25
30
35
-200
-100
0
100
200
residuals


Histogram of residuals
Frequency
0
5000
10000
15000
-200
-100
0
100
200
residuals

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

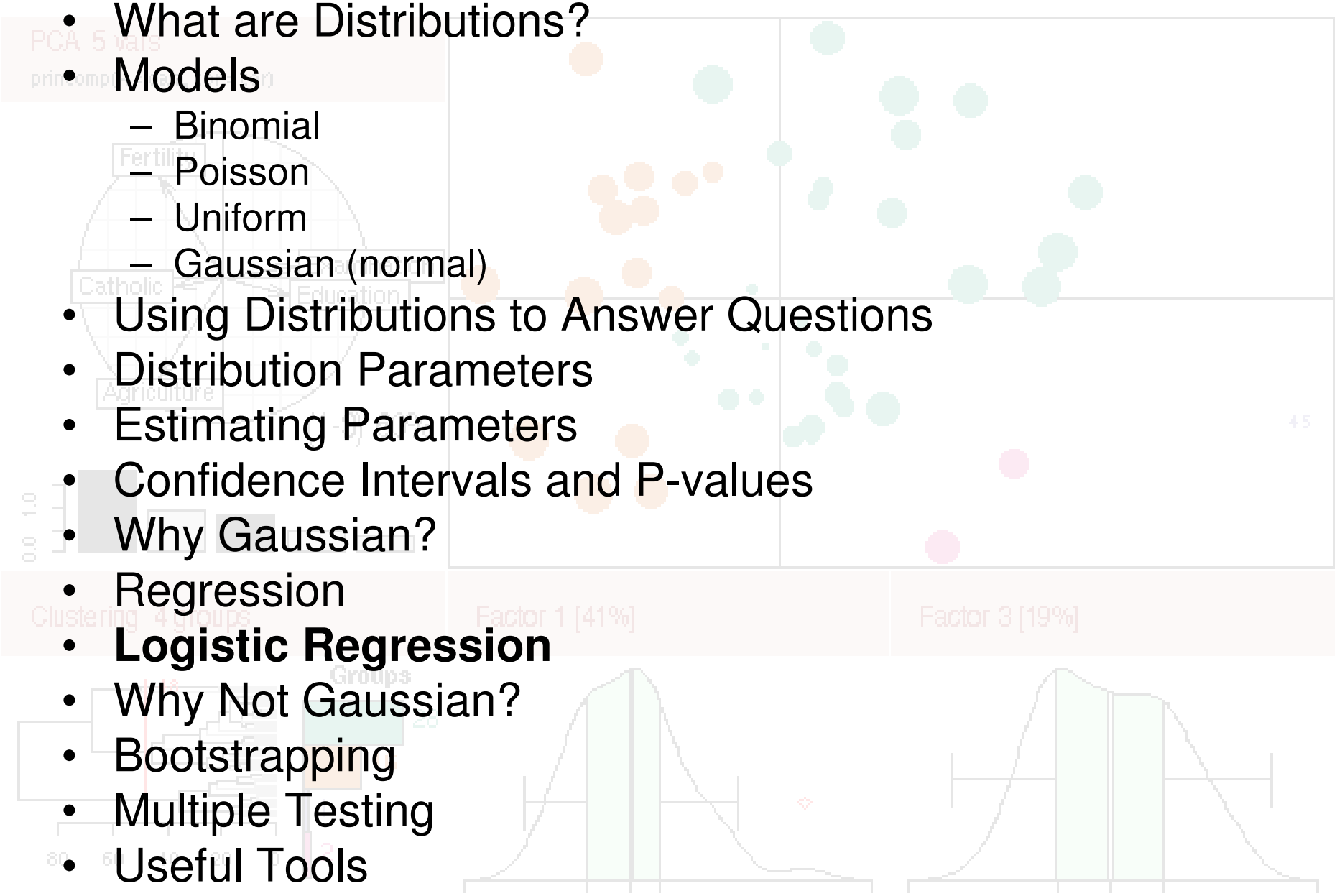
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_0 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$$

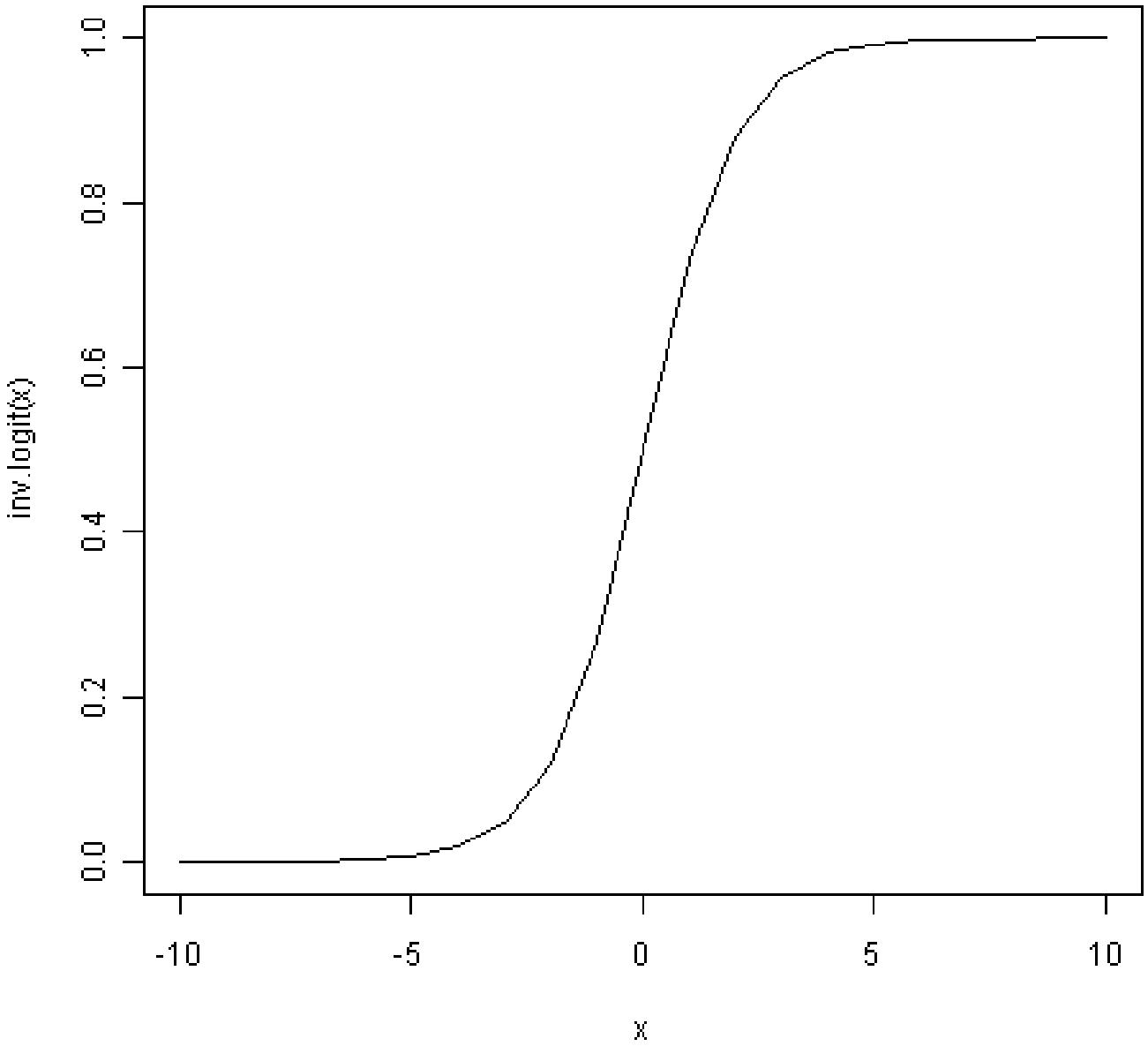
- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- Distribution Parameters
- Estimating Parameters
- Confidence Intervals and P-values
- Why Gaussian?
- Regression
- **Logistic Regression**
- Why Not Gaussian?
- Bootstrapping
- Multiple Testing
- Useful Tools



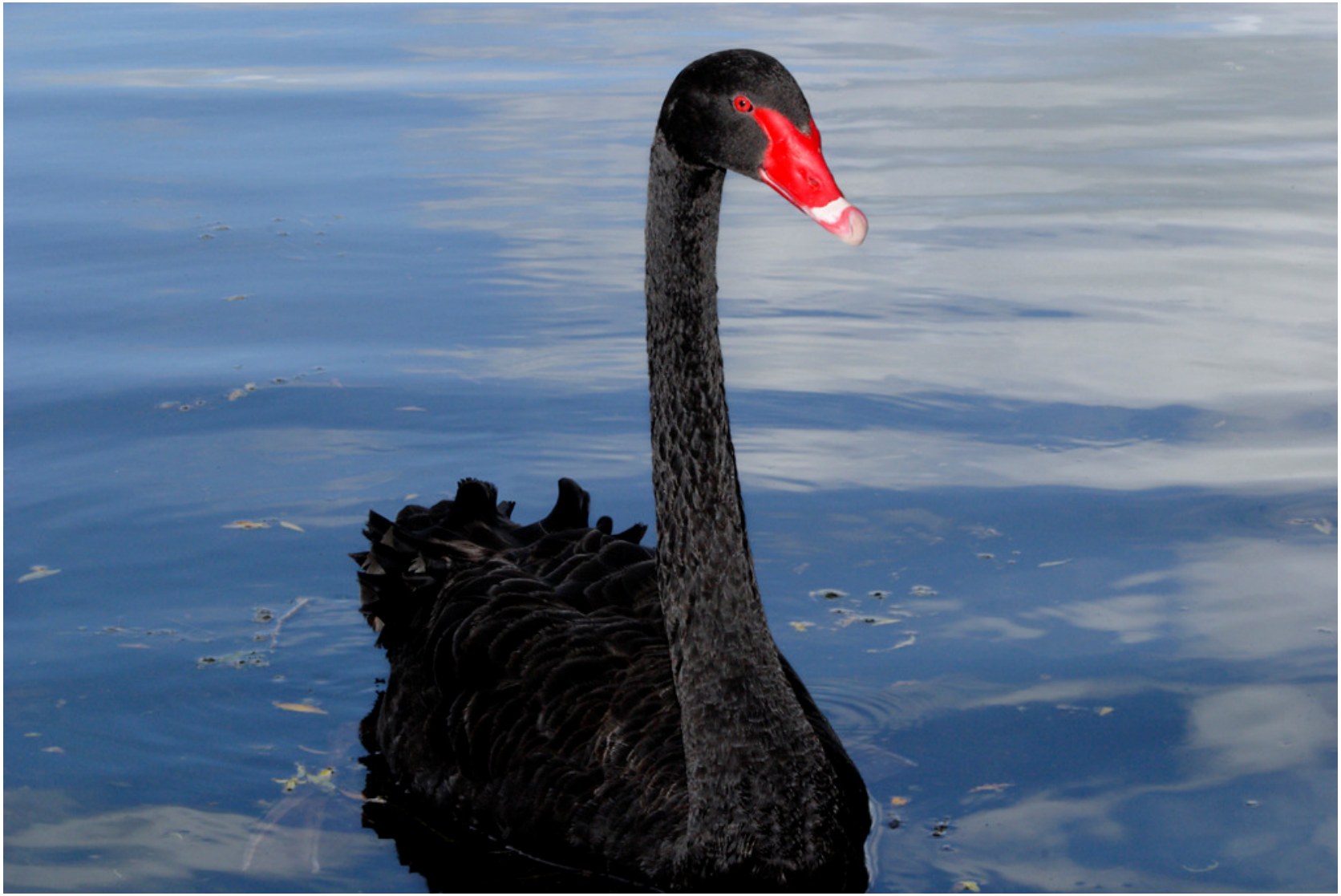
$$Y_i \sim Bin(inv.logit(\beta_0 + \beta_1 X_i), n = 1)$$

$$Logit(P_i) = \beta_0 + \beta_1 X_i$$

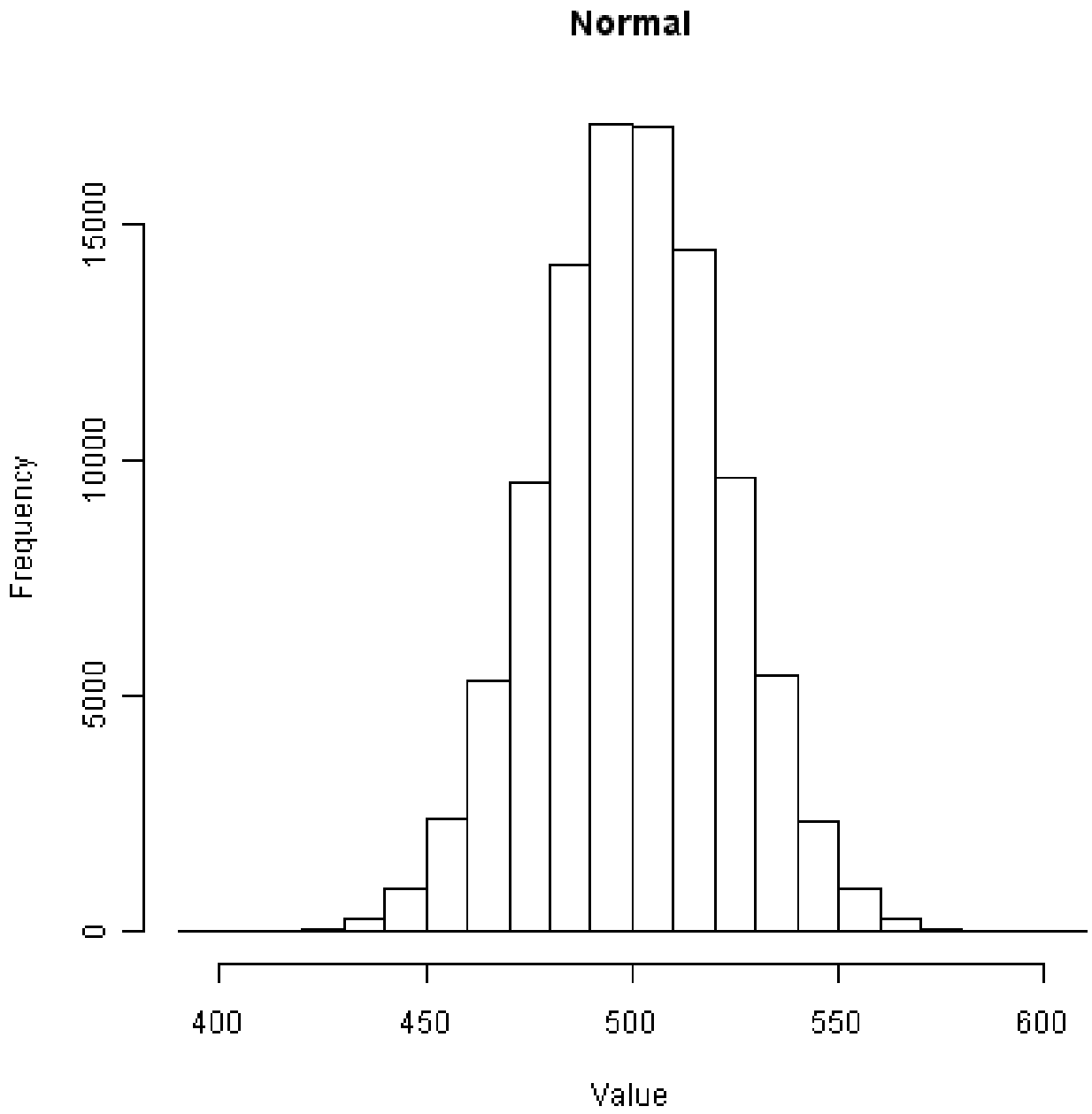
$$Y_i \sim Bin(P_i, n = 1)$$

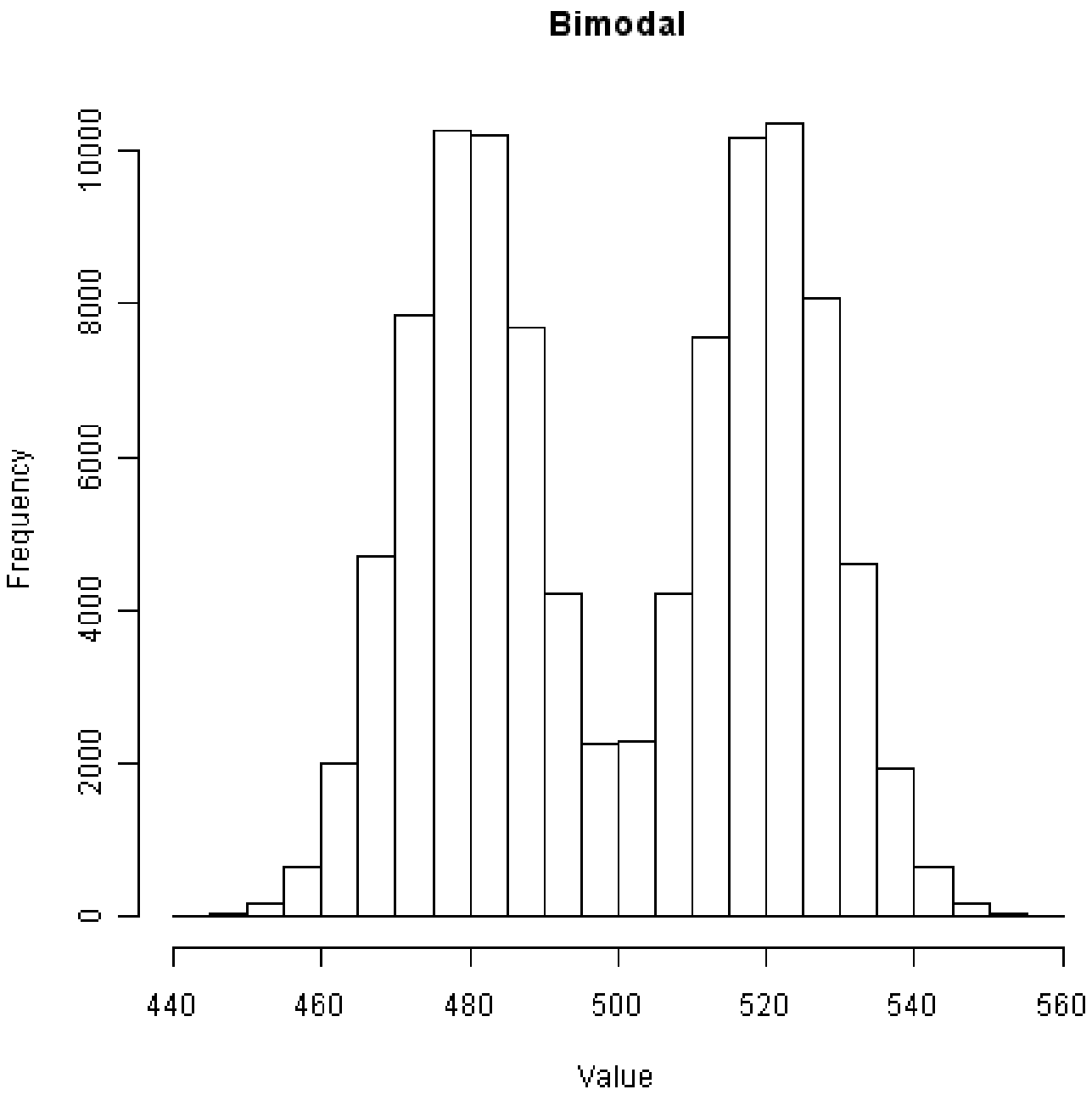

inv.logit(x)
0.0
0.2
0.4
0.6
0.8
1.0
-10
-5
0
5
10
x

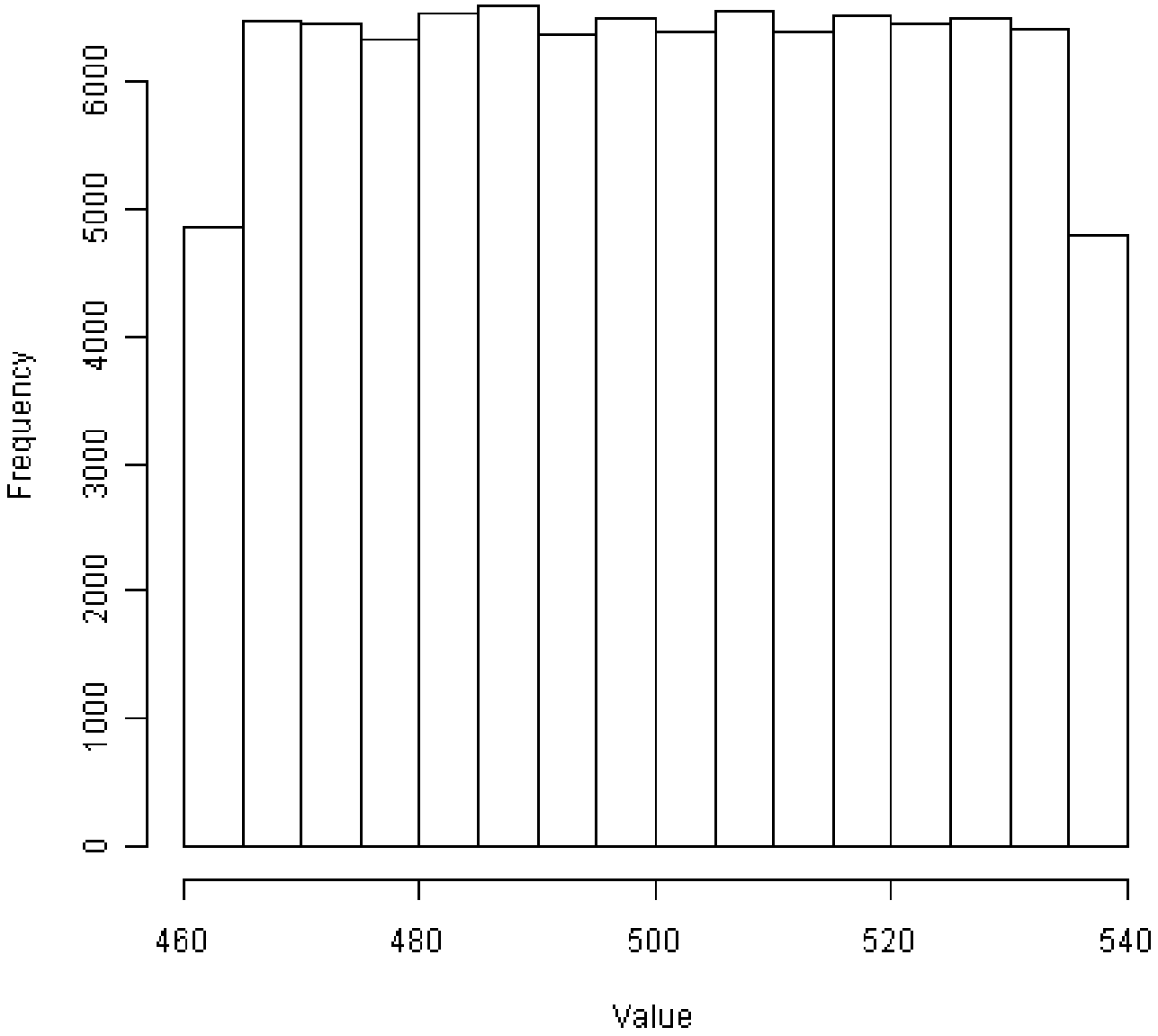
- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- Distribution Parameters
- Estimating Parameters
- Confidence Intervals and P-values
- Why Gaussian?
- Regression
- Logistic Regression
- **Why Not Gaussian?**
- Bootstrapping
- Multiple Testing
- Useful Tools

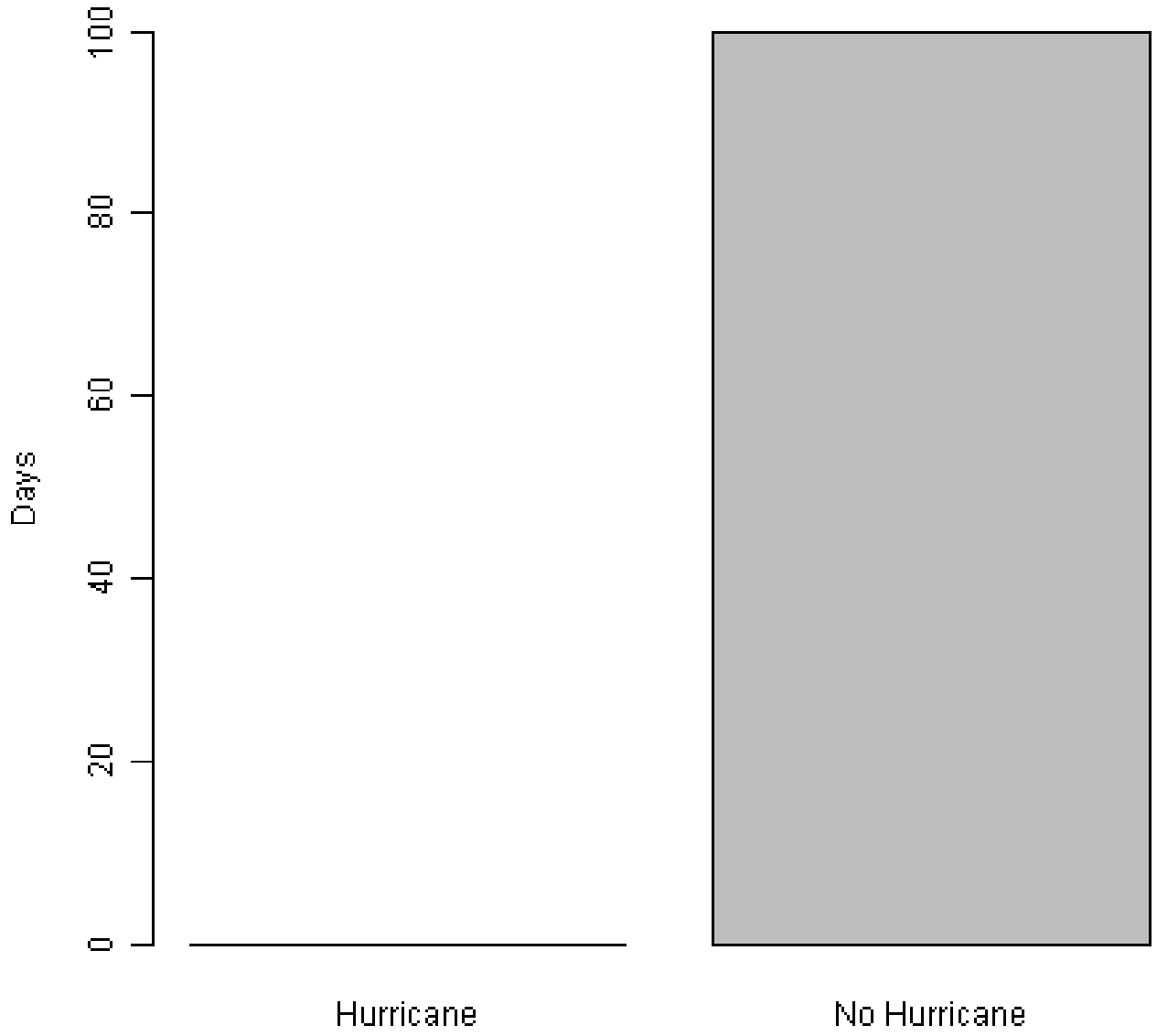


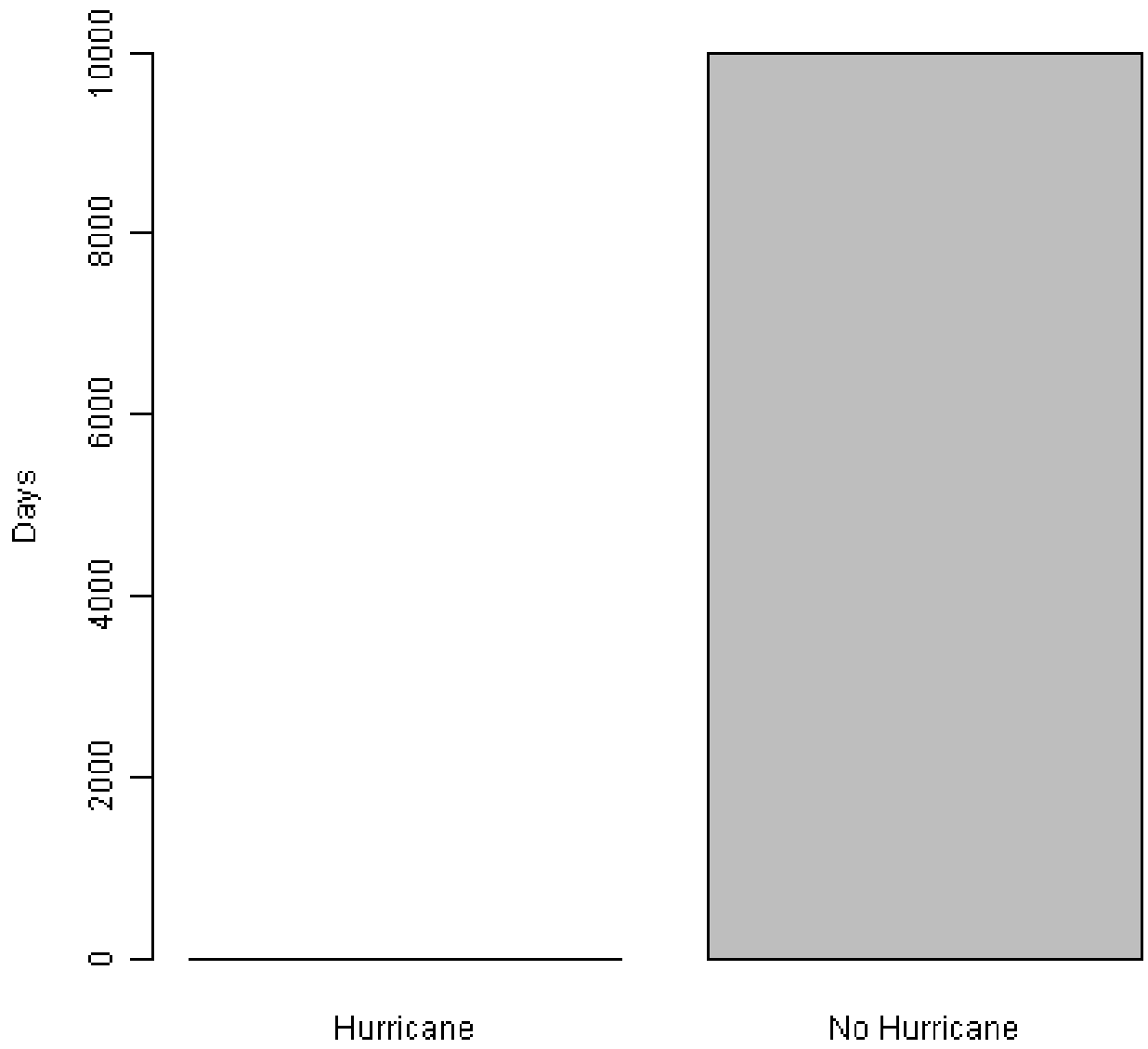
Creative Commons licensed, from Alan Vernon on flickr


Normal
Frequency
0
5000
10000
15000
400
450
500
550
600
Value


Bimodal
Frequency
0
2000
4000
6000
8000
10000
440
460
480
500
520
540
560
Value


Uniform
Frequency
0
1000
2000
3000
4000
5000
6000
460
480
500
520
540
Value

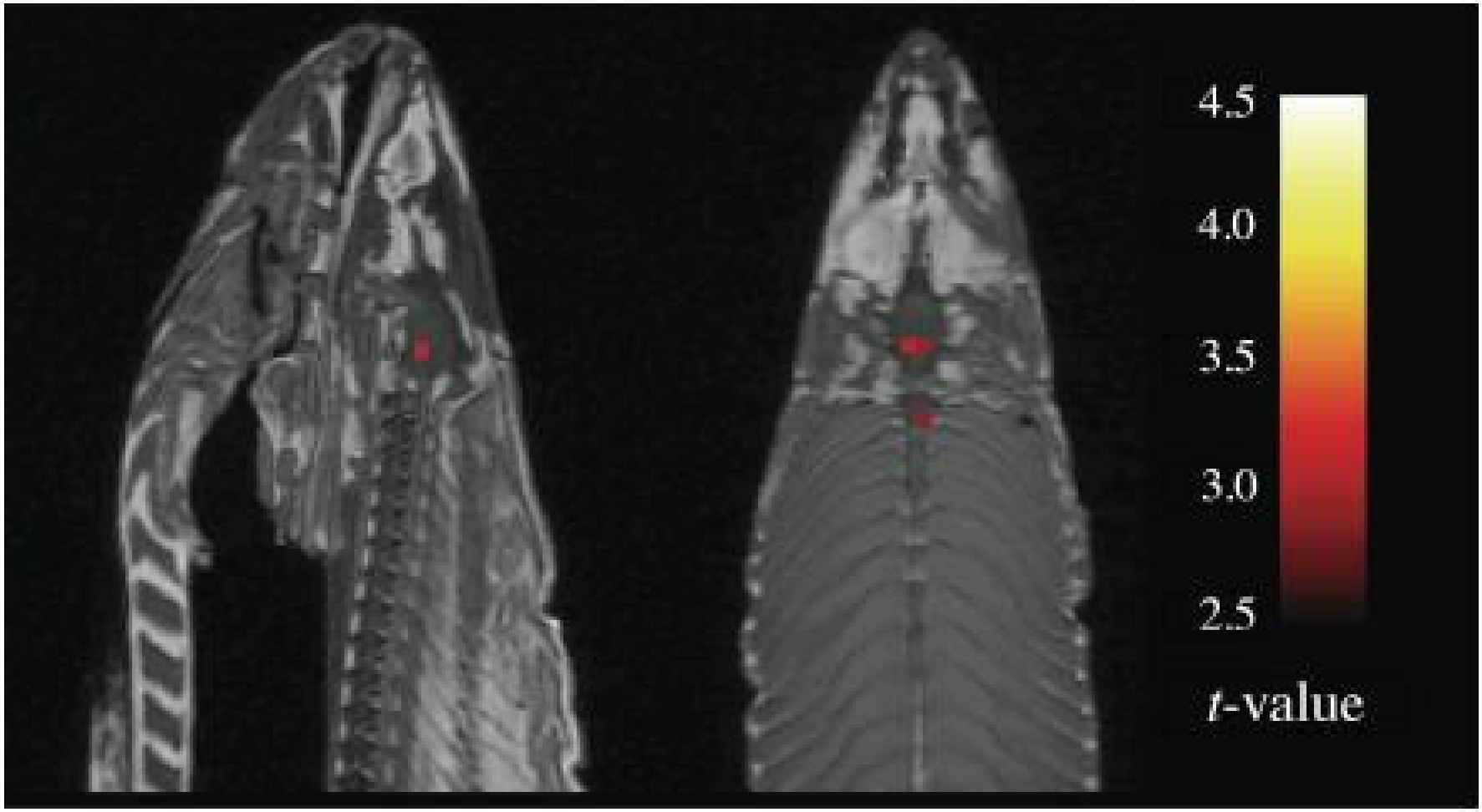

Days
0
20
40
60
80
100
Hurricane
No Hurricane


Days
0
2000
4000
6000
8000
10000
Hurricane
No Hurricane

- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- Distribution Parameters
- Estimating Parameters
- Confidence Intervals and P-values
- Why Gaussian?
- Regression
- Logistic Regression
- Why Not Gaussian?
- **Bootstrapping**
- Multiple Testing
- Useful Tools

- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- Distribution Parameters
- Estimating Parameters
- Confidence Intervals and P-values
- Why Gaussian?
- Regression
- Logistic Regression
- Why Not Gaussian?
- Bootstrapping
- **Multiple Testing**
- Useful Tools

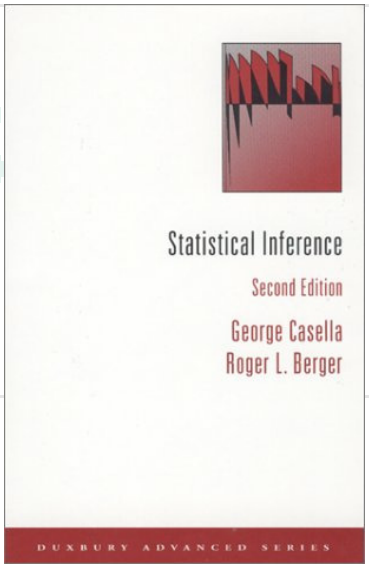
Creative Commons licensed, from CharlotteKinzie on flickr


4.5
4.0
3.5
3.0
2.5
t-value

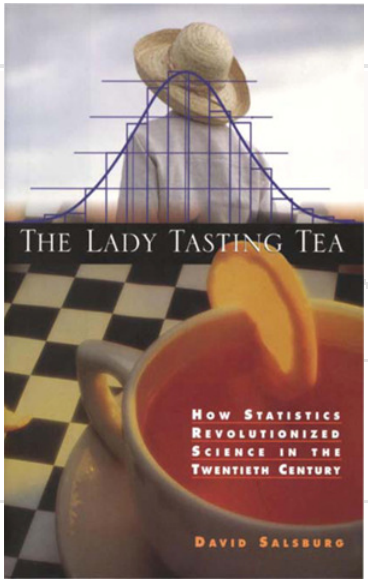
Probability of False Detection (per test)	Number of Tests	Probability of False Detection (in at least 1 test)
0.01	1	0.01
0.01	10	0.10
0.01	100	0.63
0.05	1	0.05
0.05	10	0.40
0.05	100	0.99

- What are Distributions?
- Models
  - Binomial
  - Poisson
  - Uniform
  - Gaussian (normal)
- Using Distributions to Answer Questions
- Distribution Parameters
- Estimating Parameters
- Confidence Intervals and P-values
- Why Gaussian?
- Regression
- Logistic Regression
- Why Not Gaussian?
- Bootstrapping
- Multiple Testing
- **Useful Tools**

- R: http://cran.r-project.org/
- Casella & Berger, *Statistical Inference*



- *Wikipedia:* *http://en.wikipedia.org/wiki/List_of_probability_distributions*
- Salsburg, *The Lady Tasting Tea*



- *thomas.lotze@thomaslotze.com*

1. Everything is a Distribution
2. Many Kinds of “Random” (many Distributions)
3. Estimated Parameters are Random
   They have Distributions!
4. Statistical Decisions come from Distribution Estimates
5. Be Skeptical of Normality
   Mean and Variance are not sufficient!
6. Be Skeptical of Multiple Testing

- Practical Take-home
  - Normality test
  - T-test
  - Wilcoxon rank-sum test
- Other Distributions
  - Student's T
  - F
  - Lognormal
  - Geometric
  - Levy
  - Weibull
  - Benford
- Goodness of fit
  - Chi-squared test
  - Q-Q plots
- Distribution Connections
- Multivariate Distributions
- Bias/Variance Tradeoff

- Nonparametric Distributions
- Model Comparison: Parameters and Fit (AIC, BIC)
- Bayesian Statistics
- Bayes' Law
- Cognitive Biases
- Time series
- Counterintuitive Probability
  - Monty Haul
  - Two Aces
  - Poisson Waiting Times
- Markov Chain Monte Carlo
- Extreme Value Statistics